Sensitivity Analyses in Empirical Studies Plagued with Missing Data

Sensitivity Analyses in Empirical Studies Plagued with Missing Data

A dissertation presented

by

Viktoriia Liublinska

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2013

Professor Donald B. Rubin **Viktoriia Liublinska**

Sensitivity Analyses in Empirical Studies Plagued with Missing Data

Abstract

Analyses of data with missing values often require assumptions about missingness mechanisms that cannot be assessed empirically, highlighting the need for sensitivity analyses. However, universal recommendations for reporting missing data and conducting sensitivity analyses in empirical studies are scarce. Both steps are often neglected by practitioners due to the lack of clear guidelines for summarizing missing data and systematic explorations of alternative assumptions, as well as the typical attendant complexity of missing not at random (MNAR) models.

We propose graphical displays that help visualize and systematize the results of sensitivity analyses, building upon the idea of "tipping-point" analysis for experiments with dichotomous treatment. The resulting "enhanced tipping-point displays" (ETP) are convenient summaries of conclusions drawn from using different modeling assumptions about the missingness mechanisms, applicable to a broad range of outcome distributions. We also describe a systematic way of exploring MNAR models using ETP displays, based on a pattern-mixture factorization of the outcome distribution, and present a set of sensitivity parameters that arises naturally from such a factorization. The primary goal of the displays is to make formal sensitivity analyses more comprehensible to practitioners, thereby helping them assess the robustness of experiments' conclusions. We also present an example of a recent use of ETP displays

in a medical device clinical trial, which helped lead to FDA approval.

The last part of the dissertation demonstrates another method of sensitivity analysis in the same clinical trial. The trial is complicated by missingness in outcomes "due to death", and we address this issue by employing Rubin Causal Model and principal stratification. We propose an improved method to estimate the joint posterior distribution of estimands of interest using a Hamiltonian Monte Carlo algorithm and demonstrate its superiority for this problem to the standard Metropolis-Hastings algorithm.

The proposed methods of sensitivity analyses provide new collections of useful tools for the analysis of data sets plagued with missing values.

Contents

Acknowledgments

I would like to express my sincere appreciation to my principal advisor and mentor, Professor Donald B. Rubin, for the guidance and advice that he has given me throughout my PhD program. During our extended conversations I learned a lot about the art of balancing rigor and pragmatism in real-world problems, guided by statistical intuition. It enabled me to grow and mature as a statistician.

I also thank Roee Gutman for helpful discussions and for assisting in the implementation of the data imputation procedure, Soteira, Inc. for permission to use their data, and Arman Sabbaghi for the assistance with the implementation of the HMC algorithm. I am very grateful to Dr. Gregory Campbell for pointing us to the original publication on related tipping-point analyses, and for serving on my dissertation committee and providing insightful comments.

My gratitude goes out to Professor Xiao-Li Meng for his mentorship, endless enthusiasm and continuous supply of innovative ideas and new opportunities for his students. I enjoyed being part of the Happy Team and I took away a lot from this experience. I thank Professor Carl Morris for advising me during the earlier years of my PhD program and helping me refine my research skills, and Professor Joe Blitzstein for being an incredible pedagogy mentor. I thank the entire Department of Statistics for creating an welcoming, but challenging, atmosphere that helped me develop as a researcher and as a teacher.

Lastly, I would like to give thanks to my wonderful fiancé, Yves Rene Chretien, for supporting and inspiring me on my PhD journey.

Chapter 1

Missing Data in Empirical Studies

The best solution to handle missing data is to have none.

- R.A. Fisher

1.1 Missing Data Mechanisms

When Ronald A. Fisher and Jerzy Neyman were laying the foundation of modern Statistics at the beginning of the 20'th century, the problem of missing data naturally emerged from the applied work conducted by researchers in various fields. One of the first published methods to account for missing observations was developed for field experiments in Allan and Wishart (1930). It was later generalized in Yates (1933) and is now regarded as a classical method of handling missing data using ANOVA (Little and Rubin 2002, p. 28). M'Kendrick (1925) studied numerous medical data and, when calculating the infection rate in the population, proposed a method to solve the issue with unobserved exposure indicator. His approach was later recognized to be a

special case of one of the most widely used methods of handling data with missing values, the EM algorithm (Dempster et al. 1977; Meng 1997). Wilks (1932) was the first to formally employ method of maximum likelihood, introduced by R. A. Fisher a decade earlier, to provide inference on population parameters in a bivariate normal setting with missing observations.

R. A. Fisher, the greatest statistician of his time, was undoubtedly right by implying that the most effort should be devoted to prevention of missing data. However, it is almost inevitable that the issue will come up in applications, and most data analysis procedures are not designed to handle it. The problem of non-random attrition and nonresponse[1] in survey research as well as missing data in randomized experiments has been widely addressed in the literature (Rubin 1987; Schafer 1997; Little and Rubin 2002; Allison 2001; McKnight 2007). Nevertheless, up until a half a century later, missing values in applied work were handled primarily by editing or case deletion (Schafer and Graham 2002). Only with the formalization of a framework of inference from incomplete data developed in Rubin (1976), the research of methods to handle missing data began to gain momentum.

Missing data pose a major problem for experiments as well as observational studies. If proper randomization was performed, the presence of missing data jeopardizes the original balance of the design and may lead to invalid inferences if not handled properly. Observational studies also suffer from missing data in covariates that are believed to be important in predicting the treatment and outcome, or in the outcome itself, especially if the missing data mechanism is unknown, which is usually

[1]Here and throughout the article we assume *item nonresponse*, implying that some information about each missing unit is available.

the case. Improper analysis of incomplete data can result in reduced statistical power, decreased generalizability of findings, and biased parameter estimates.

Here we adopt a standard approach to define and classify missing data. A value is considered *missing* if it is potentially observable and meaningful for analysis, although not available in the data set at hand. With N units in the dataset, let $\boldsymbol{X} = (x_{ik}) = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K)$ be the NxK matrix of baseline variables (*covariates*, or *predictors*), and let $\boldsymbol{Y} = (y_{ij}) = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_J)$ be the NxJ matrix of outcome measures (or *dependent variables*). It is important to distinguish missingness in baseline predictors and in outcomes because it may have to be handled differently (Little 1992; Moons et al. 2006; Newgard and Haukoos 2008).

We define a matrix of missingness indicators for the outcomes, $\boldsymbol{D}_Y = (d_{ij})$, such that $d_{ij} = 1$ if unit i is missing the jth outcome. Analogously, a matrix of missingness indicators for the baseline variables is defined as $\boldsymbol{D}_X = (d_{ik})$, and we let $\boldsymbol{D} = (\boldsymbol{D}_X, \boldsymbol{D}_Y)$. The paramount idea introduced in Rubin (1976) suggests that we need to regard the d_{ij} as random variables, and offers a straightforward way to define missing data mechanisms through distributions on the d_{ij}.

Let a set $\boldsymbol{Y}_{obs} = \{y_{ij} \mid d_{ij} = 0\}$ contain the observed values among the outcomes, and a set $\boldsymbol{Y}_{mis}$ contain the missing elements of the matrix $\boldsymbol{Y}$, such that $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$; note that $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ are *not* matrices, but rather collections of elements of the matrix $\boldsymbol{Y}$, where, formally, the sets *obs* and *mis* are functions of $\boldsymbol{D}_Y$. Analogous sets can be defined for the matrix of baseline variables, $\boldsymbol{X} = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis})$. Also, let $f(\boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\phi})$ be the conditional distribution of missingness indicators given all data values, observed and missing, and unknown vector-parameter $\boldsymbol{\phi}$.

The missingness mechanism is called *missing completely at random* (MCAR) if, for each possible value of $\boldsymbol{\phi}$,

$$f(\boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\phi}) = f(\boldsymbol{D} \mid \boldsymbol{\phi}) \text{ for all } \boldsymbol{D}, \boldsymbol{X}, \text{ and } \boldsymbol{Y}.$$

In other words, in a simple case with one vector of outcomes and no predictors ($K = 1$ and $J = 0$), missing values can be viewed as randomly deleted. However, in higher dimensions, $K > 1$ or $J > 0$, or both, it is allowed for the missingness indicators to interact, though independently from the data.

It is rarely the case that the MCAR assumption holds in practice. One scenario where the MCAR assumption is plausible is when the data were deliberately not collected, or *missing by design* (Rubin 1987). The less restrictive *missing at random* (MAR) assumption holds if, for each possible value of $\boldsymbol{\phi}$,

$$f(\boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\phi}) = f(\boldsymbol{D} \mid \boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs}; \boldsymbol{\phi}) \text{ for the observed } \boldsymbol{D}, \boldsymbol{X}_{obs}, \text{ and } \boldsymbol{Y}_{obs},$$

and for all $\boldsymbol{X}_{mis}$ and $\boldsymbol{Y}_{mis}$, i.e., if the distribution of missingness indicators depends only on the observed covariate and outcome values. Although this is how MAR assumption was defined originally in Rubin (1976) for the purpose of Bayesian or direct-likelihood inference, it is sometimes mistakenly employed for sampling distribution (or frequentist) inference based on a large-sample theory, e.g., constructing confidence intervals (see Heitjan and Basu 1996).

A stochastic generalization of MAR that allows to utilize frequentist inference, called a "MAR mechanism" in Little and Rubin (1987), was formally called *missing*

always at random (MAAR) in Mealli and Rubin (2013), which holds if the following is true:

$$f(\boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\phi}) = f(\boldsymbol{D} \mid \boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs}; \boldsymbol{\phi}) \text{ for all } \boldsymbol{D}, \boldsymbol{X} \text{ and } \boldsymbol{Y},$$

and for each possible value of $\boldsymbol{\phi}$. In other words, the missingness should depend on the observed data only, and it should hold for *all* realizations of the missing-data pattern $\boldsymbol{D}$ and random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, not just for the observed ones. This condition requires analysts to consider a hypothetical missingness mechanism even in cases when all values in the data were observed, as long as some of them could potentially have been missing. However, MAR would hold if, for units with covariate or outcome missingness depending on the underlying values, all values were observed in a current realization. In addition, Little (1995) introduced the term *covariate-dependent* (CD) missingness for situations with no missingness in predictors ($\boldsymbol{X} = \boldsymbol{X}_{obs}$). CD missingness is a special case of MAAR when the missingness mechanism depends only on predictors and not on the outcomes, i.e.,

$$f(\boldsymbol{D} \mid \boldsymbol{X}_{obs}, \boldsymbol{Y}; \boldsymbol{\phi}) = f(\boldsymbol{D} \mid \boldsymbol{X}_{obs}; \boldsymbol{\phi}) \text{ for all } \boldsymbol{D}, \boldsymbol{X}_{obs} \text{ and } \boldsymbol{Y},$$

for each possible value of $\boldsymbol{\phi}$. In fact, this assumption is the one most commonly used in practice, although many studies erroneously report using MAR assumption.

Assume that the joint distribution of outcomes $\boldsymbol{Y}$ and predictors $\boldsymbol{X}$ has a probability model $f(\boldsymbol{Y}, \boldsymbol{X} \mid \boldsymbol{\theta})$, governed by unknown vector-parameter $\boldsymbol{\theta}$, and suppose we are interested in estimating $\boldsymbol{\theta}$. The missing data are said to be *ignorable* for the purpose of likelihood-based inference for $\boldsymbol{\theta}$ if MAR is satisfied and parameters $\boldsymbol{\phi}$ carry

no information about $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are *distinct*[2], Rubin 1976; Little and Rubin 2002). The term "ignorable" comes from the fact that $f(\boldsymbol{D} \mid \boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs}; \boldsymbol{\phi})$ may be "ignored" (or dropped) from the likelihood without altering the likelihood function (or posterior distribution) of $\boldsymbol{\theta}$.

If either the distinctness of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ or MAR is not met, missing data are considered *nonignorable*. Violation of distinctness is less consequential than violation of MAR, because the likelihood-based inference will still produce consistent, although generally inefficient, estimates, whereas, violation of MAR is often critical (see Appendix A.1 for an example of nonignorable MAR mechanism). If the missingness mechanism does not satisfy MAR, it is regarded as *missing not at random* (MNAR). Analysis of data with MNAR missingness requires specifying a full-data likelihood $f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$, including a model for the missingness mechanism $f(\boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\phi})$, in order to produce a generally valid likelihood-based inference. In practice, these models require making assumptions about the distribution of missing values that often cannot be assessed empirically, and, therefore, the obtained results should be subjected to sensitivity analyses.

1.2 Parameter Estimation with Incomplete Data

The most basic approach to handle data with missing values is complete-case (listwise deletion) or available-case (pairwise deletion) analysis. However, quite a few research articles were written about the shortcomings of this approach (e.g., Rubin

[2]I.e., are in disjoint parameter spaces; the concept can be extended to Bayesian inference by requiring two vector-parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, to be a priori independent.

1987; Greenland and Finkle 1995; Schafer and Graham 2002; van der Heijden et al. 2006; Carpenter and Kenward 2008; Liublinska and Rubin 2012).

Many superior methods were developed during a second half of the 20th century. One class of methods utilizes the idea introduced by Horvitz and Thompson (1952), which suggests weighting responses by inverse-probability of observation to produce unbiased estimate of population averages (**?**Little and Rubin 2002, section 3.3). One can think of this procedure as a reconstruction of the population of interest, with each weight corresponding to an approximate number of units in the population that the observed response represents.

The Horvitz-Thompson estimator was originally proposed for analysis of surveys with sampling weights set in advance and the estimator is most efficient when the true weights are known. However, in observational studies it is largely impossible to know the missingness mechanism exactly and weighting methods require modeling response probabilities using available covariates. Estimates based on weighting responses by the estimated propensity to respond can be very unstable; they rely heavily on the validity of the proposed propensity model. In addition, if very few respondents are similar to nonrespondents, they would have a disproportionately large effect on the estimate, resulting in large uncertainty bounds. Inference from this class of methods is mainly focused on marginal population characteristics, e.g., average response, although it can be extended to consistently estimate parameters from the conditional distribution $f(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta})$ (Robins et al. 1994, 1995). In addition, by incorporating a model for the response itself, *doubly-robust* estimators can be constructed (see **?**, for an extensive review).

A few model-based approaches to draw valid inferences in the presence of missing data have been developed over the last several decades. One is based on specifying full (or observed) likelihood of the data and performing MLE estimation using various maximization methods, including expectation-maximization (EM) (Dempster et al. 1977), Newton-Raphson, or scoring algorithms. A Bayesian analog of this estimation approach extends the model by adding a prior component for parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, $p(\boldsymbol{\theta}, \boldsymbol{\phi})$, and estimating their joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$ (Tanner and Wong 1987).

The full-likelihood approach is quite complex analytically and computationally; it requires joint parametric modeling of the data-generating process and, sometimes, the missing data mechanism too. In the EM algorithm, the M-step may be hard to formulate and the convergence to the maximum can be particularly slow if the fraction of missing information is large. However, if the model is specified correctly, MLE estimate has attractive large-sample properties, including consistency, asymptotic normality and asymptotic efficiency.

Another class of methods recommends imputing each missing response. Then, any quantity of interest from the conditional (and marginal) distribution of the response can be easily obtained from a resulting rectangular dataset. Imputation methods can be classified into model-based and hot-deck. The former class utilizes the relationship between the response and available covariates. Consequently, the estimates strongly depend on the accuracy of the model. The hot-deck class offers procedures to match the respondents and nonrespondents and impute the missing responses by drawing from a "donor pool" of units with observed responses (e.g., exact matching, predictive

mean matching (Rubin 1986; Little 1988a), propensity score matching (Little 1986), etc.). Note that all these methods require substantial overlap between respondent's and nonrespondent's covariates, a problem that we cover in details in Section 1.3.2 below.

Some imputation methods involve producing single imputation for each missing value, e.g., mean substitution, regression substitution, worst-case substitution, last observation carried forward (LOCF). Although there are settings where these methods will result in valid inferences, they require strict assumptions that are, often, unrealistic (Little and Rubin 1987, 2002; Rubin and Schenker 1991; Little 1992; Schafer 1997, 1999; Donders et al. 2006).

A more general imputation approach that has been gaining momentum over the last decade is multiple imputation (MI, Rubin 1987), i.e., creating multiple completed datasets by imputing missing values from their posterior predictive distribution $f(\boldsymbol{Y}_{mis}, \boldsymbol{X}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}, \boldsymbol{\phi})$. If the data $(\boldsymbol{Y}, \boldsymbol{X})$ are jointly normally distributed, the posterior predictive distribution for missing values is easily derived. However, often it is too difficult to obtain $f(\boldsymbol{Y}_{mis}, \boldsymbol{X}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}, \boldsymbol{\phi})$ in a closed traceable form and a convenient algorithm was developed to provide a way to approximate the posterior predictive distribution for missing values without the need to put a model on a full joint distribution of $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{D}$. The algorithm consists of iterating a sequence of univariate regression models for imputation (Raghunathan et al. 2002), also known as "multivariate imputation by chained equations" (MICE, Rubin 2003; Buuren van and Groothuis-Oudshoorn 2011; Buuren 2012). It involves performing univariate imputations iteratively, each time fitting a model to a variable with missing values,

conditional on all others included in the analysis. A variable with newly imputed values is conditioned on in subsequent iterations, and the procedure cycles though all variables with missing values until the convergence of the sampling distribution of imputed variables is achieved.

The advantage of MI is that it enables practitioners to use widely-available complete-data methods on each imputed dataset separately and incorporate the uncertainty due to the presence of missing data by pooling the results using Rubin's Combining Rules (Rubin 2004). More important, this method is very suitable for performing sensitivity analyses, because one can use multiple models to generate imputations and compare conclusions across the models. In Section 3.3 we show how MI can be utilized to explore the consequences of alternative assumptions about the missing data mechanism.

Despite a plethora of available methods to produce valid inference for incomplete data, to this day, very few empirical studies acknowledge this issues and, even less, handle it properly. As we show next, there are no agreed upon guidelines on reporting the amount and the characteristics of missing data in a study, and the decision to report them is usually made at the practitioner's discretion. As a result, even when study reports indicate that some data are missing, most of them do not discuss assumptions that were made regarding the missing data or missingness mechanisms, nor do they include any sensitivity analyses.

1.3 Standards of Missing Data Reporting

A major breakthrough was made in the ways missing data are handled in empirical studies due to the effort of many outstanding statisticians to study and explain the extent of the issue to practitioners. It is now a common knowledge that reporting the presence of missing data is necessary, although still seldom done in practice, and an increasing number of studies attempt to employ the methods described in Section 1.2 to account for the missingness. However, there are very few explicit reporting guidelines, approved and agreed upon in statistical community, available for analysts who work on studies with missing data. This shortcoming results in lack of structure in reporting practices observed throughout the literature. The danger is that haphazard and fragmented description of missing data may result in a false assurance in study's conclusion.

Several revealing surveys of articles in empirical research journals were conducted in recent years. Their objective was to study missing data prevalence, reporting, and handling practices, and their conclusions were worrisome, but promising. For instance, White et al. (2011) reviewed randomized controlled trials published in major medical journals in 2001. Out of 71 trials that were surveyed, 89% reported having partly missing outcome data. Among those, 65% performed complete case analysis, most of the rest performed single imputation, and only 21% conducted some sensitivity analysis. Klebanoff and Cole (2008) and Mackinnon (2010) focused on studies that used MI and concluded that, although MI is becoming more common in medical studies, clear guidelines of reporting MI procedure should be developed. Other surveys of clinical studies (Burton and Altman 2004; Aylward et al. 2010) reported

similar observations.

The situation is more alarming in social sciences. Bodner (2006) reported that, in a random sample ($N = 181$) of empirical studies taken from almost 36,000 articles identified in PsycINFO database (that contains studies in social and behavioral science research) in 1999, two-thirds either did not have missing data or failed to report them completely. Among the rest, only half explicitly discussed missing data in the text, and a vast majority (97%) did not account for them in any way (i.e., used either complete-case or available-case analysis). See also Jelicic et al. (2009) for a review of similar studies.

Another article Peugh and Enders (2004) provided a much larger methodological review of missing-data reporting practices in 23 applied educational and psychological journals published in 1999 and 2003 (around $1,500$ articles in total). The findings for 1999 were consistent with the ones reported in Bodner (2006), i.e., "*details concerning missing data were seldom reported*" and methods used to address the issue were rudimentary. Conclusions from articles published in 2003 were more optimistic: half of studies had some indication of the presence of missing data and most of those explicitly discussed the problem in the text. The authors attribute this improvement to a previously published report by the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson 1999), which provided many important guidelines on current practices of data analysis. However, a thorough review of the report identified only one sentence that touches upon the missing data issue: "*Before presenting results, report complications, protocol violations, and other unanticipated events in data collection. These include missing data, attrition, and nonresponse*".

Although the message is important, more details are needed.

The most complete and up-to-date report on the issue of prevention and treatment of missing data (focusing on clinical trials) was assembled by the National Academy of Sciences at the request of the U.S. FDA (NRC-Panel 2010). Although the authors gave full and detailed account of modern techniques for handling missing data, surprisingly little attention was devoted to standardizing their reporting. Evidently, it reflects lack of research in this area, as authors themselves admitted the need for more standardized documentation and analysis of the reasons for missing data (NRC-Panel 2010, p.112).

Current industry standards of reporting randomized trials (Schulz et al. 2010) require researchers to disclose the number of excluded participants after randomization and the reasons for the exclusion, without any more details. Below we demonstrate that further analysis of characteristics of the study participants with missing observations, as well as the exact time of their dropout, may be crucial in assessing the appropriateness of chosen missing-data techniques, including checking if the required assumptions are scientifically justifiable.

A unique report that provides a more rigorous treatment of the process of missing data exploration was issued by the European Medicines Evaluation Agency (EMEA, CHMP 2010). This report emphasizes the importance of a thorough discussion of the amount, reasons, timing, and pattern of missing data, and possible implications of having it. Although many crucial reporting elements were emphasized, the practical advice was still sparse. Next we discuss some elements mentioned in guidelines issued by EMEA, expand upon them and present formal and graphical methods that can be

used to report missing data in empirical studies.

1.3.1 Important Missing Data Summaries

Every report that documents an empirical study usually contains a section on data description. It is crucial for the information about missing observations to become an essential part of this section. Below we list several recommendations on the most important components that must not be omitted from the report, with brief reasoning.

Missingness rates. The basic statistics that provide initial idea of the amount of missingness are proportions of missing values observed in each variable (possibly by treatment arm, if applicable). In fact, missing data indicators may be considered as additional outcomes, especially if missingness rates are substantially different across treatment arms. For example, in a clinical trial setting, they may be a proxy for patient's tolerance levels and treatment preference. Rates that are much higher than the difference in success rates between treatment groups may lead to a rejection of the trial by the FDA (NRC-Panel 2010).

Reasons for missingness. This information should play a major role in assessing and justifying assumptions about missing data employed in a study. Attention should be paid to reasons that relate missingness mechanism to the unobserved data. For example, survey participants refusing to respond due to the sensitive nature of the question (e.g., if some answer choices are controversial), patient dropout initiated by an adverse side effect of a new drug, or measurement censoring due to malfunction of a measuring device. These are just a few situations where the assumption of ignorable missing data would be inappropriate.

Note that the information about reasons for the missingness may help *rule out* MCAR or MAR assumptions, however it will not be sufficient to validate either of them. In general, it is difficult to fully justify ignorability assumption using available data only (Rubin 1976). It was shown that models based on MAR and MNAR assumptions could have comparable fits to the observed data but substantially different predictions for the missing part (Molenberghs et al. 1999). The decision for or against ignorability should be made after acquiring a sufficient understanding of the scientific aspect of the problem and consulting with collaborators acquainted with the field.

Studying **times of dropout and reasons for it** in randomized studies also helps to understand the missing data pattern. Committee of Health and Medicinal Products (CHMP 2010) recommended ways of summarizing the pattern of drop-outs using graphical displays. In longitudinal studies, Kaplan−Meier plots can be used to compare the time to withdrawal between each treatment group, possibly grouped by reasons. The authors also emphasized that baseline and post-baseline characteristics of subjects who discontinued and who completed the trial should be compared. This important detail is being overlooked in virtually every empirical study and we discuss it in details below.

Differences in baseline characteristics. Both CHMP (2010) and Burzykowski et al. (2010) briefly mentioned the importance of checking for any differences observed between respondents and nonrespondents, and here we elaborate on it and provide some practical advice. Peugh and Enders (2004) reported small but positive shift in the prevalence of testing MCAR assumption in 2003 sample of studies compared to the 1999 one. In fact, this is the only assumption that can be tested explicitly, even

if there are no fully observed covariates for all units (Little 1988b); any substantial differences observed between respondents and nonrespondents within any treatment group immediately exclude MCAR assumption. Besides, the size of the difference alerts us about a potential extrapolation that may take place if the inference is drawn for all subjects, including the nonrespondents, especially if the procedure relies on the observed data only. In the following section we describe a set of analytic and graphical methods for comparing various characteristics of respondents and nonrespondents and reporting found imbalances.

The above list is not exhaustive and there may be other study-specific information crucial to disclose. However, it contains some of the most important components that help to choose the appropriate way to handle missing data. In addition to the initial summaries, the **detailed description of methods** used to address missingness should be included in the analysis section, **along with assumptions** that were employed, their justification and appropriate references.

Finally, every reviewed guideline especially stressed the significance of performing sensitivity analyses that assess the impact of missing data on reported estimates and conclusions. In Chapters 2 and 3 we propose a convenient model-based procedure for conducing sensitivity analyses in studies with two treatment arms, incorporating graphical displays.

1.3.2 Assessing the Overlap Between Respondents and Nonrespondents

Here we focus on one of the most informative, but rarely studied, features of units in a study with missing data, namely, the extent to which the units with and without missing data look alike. Most software packages that perform missing data imputation automatically do not alert a user if the characteristics of respondents and nonrespondents are very dissimilar. Moreover, hardly any article that discusses methods of addressing missing data issues stresses the necessity of measuring the overlap between the characteristics of missing and observed units before conducting the analysis. However, if there is little overlap then the inference will require extrapolation.

For simplicity, we assume that there is no missing data in covariates ($\boldsymbol{X} = \boldsymbol{X}_{obs}$, $\boldsymbol{D} = \boldsymbol{D}_X$) and that the outcome is univariate ($J = 1$, extensions to incomplete multivariate outcomes are discussed as well). Suppose that units are independent and exchangeable, which allows us to drop the index "i", $i = 1, \ldots, N$, and that $f(\boldsymbol{x} \mid \boldsymbol{\nu})$ is the joint probability distribution of covariates for each unit, where $\boldsymbol{\nu}$ is a vector of parameters. We showed in Section 1.1 that, by definition, MCAR assumption holds if the distribution of the missingness indicator d does not depend on any observed data, i.e., $f(d \mid \boldsymbol{x}; \boldsymbol{\nu}) = f(d \mid \boldsymbol{\nu})$. Therefore, it follows that

$$\begin{aligned} f(\boldsymbol{x} \mid d; \boldsymbol{\nu}) &= f(\boldsymbol{x}, d \mid \boldsymbol{\nu}) / f(d \mid \boldsymbol{\nu}) \\ &= f(d \mid \boldsymbol{x}; \boldsymbol{\nu}) f(\boldsymbol{x} \mid \nu) / f(d \mid \boldsymbol{\nu}) \\ &= f(\boldsymbol{x} \mid \boldsymbol{\nu}). \end{aligned}$$

Thus, the joint distribution of covariates for nonrespondents, $f(\boldsymbol{x} \mid d = 1; \boldsymbol{\nu})$, is the same as the one for respondents, $f(\boldsymbol{x} \mid d = 0; \boldsymbol{\nu})$. This fact can be used to construct a wide variety of tests and graphical summaries to verify the MCAR assumption. Moreover, even if we conclude that the MCAR assumption is not justified, the results of these tests can help us to assess the amount of extrapolation that will occur if MAR (or MNAR) assumption is used instead. For example, in Section 2.3.2 we present data from a randomized clinical trial, where some groups of subjects with missing outcomes did not resemble any of the subjects with fully observed outcomes. We used these discrepancies to conduct further sensitivity analyses of the study's conclusions. Next, we describe several analytical and graphical methods that can be used to quantify and qualify the overlap between respondents and nonrespondents.

Analytical methods. A straightforward way of comparing units with and without missing outcomes consists of comparing the distribution of fully-observed covariates one at a time as well as their two-way interactions (or any other function of $\boldsymbol{x}$). Standard tests, such as t-test for means or z-test for proportions, F-test for variances, and Kolmogorov-Smirnov test for empirical distributions, can be used in any combination, depending on the distributions of covariates under consideration.

The next step is to consider summaries based on all covariates for respondents and nonrespondents, comparing features of their joint distributions. For instance, for a subset of covariates whose joint distribution resembles a Multivariate Normal distribution we can calculate the Mahalanobis distance between the group means,

$$H^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0)' \, \hat{C}^{-1} \, (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0) \,,$$

where $\bar{\boldsymbol{x}}_0$ and $\bar{\boldsymbol{x}}_1$ are vectors of sample means of covariates for respondents and non-respondents, respectively, and $\hat{C}$ is the estimated pooled covariance matrix. Various test statistics have been developed to test if the underlying populations' means are equal (e.g, Hotelling T^2, Hotelling 1931; Cacoullos 1965a,b).

Finally, many tests were proposed to check the MCAR assumption in situations with missingness in more than one variable. Majority of them propose testing homogeneity of means and covariances among the multiple groups of data, distinguished by their missing data patterns, i.e., groups of units that have missing values in the same set of variables, (e.g., Little 1988b; Kim and Bentler 2002; Jamshidian and Jalal 2010). Several R-packages were created to facilitate testing MCAR assumption, e.g., `MissMech`, `BaylorEdPsych`.

Graphical methods. Remarkably, some ideas for comparing the subgroups of respondents and nonrespondents can be borrowed from the theory of unit-matching in causal inference (Rubin 2006b). Indeed, the procedures for checking balance between matched treated and control units have the same goal, i.e., to ensure that joint distributions of covariates are sufficiently similar.

First obvious step is to compare histogram shapes and check the difference in ranges. An effective method of assessing the balance visually, so called *Love plots*, was introduced by Thomas E. Love (Ahmed et al. 2006; D'Agostino Jr. 1998). The plots display standardized differences in average covariate values between two groups, calculated for discrete and continuous variables as follows:

$$d_c = \frac{100\,(\bar{x}_1 - \bar{x}_0)}{\sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_0^2)\,/2}},\quad d_b = \frac{100\,(\hat{p}_1 - \hat{p}_0)}{\sqrt{(\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0))\,/2}}.$$

Generally, absolute standardized differences above 10% indicate serious imbalance. These plots are utilized in the example in Section 2.3.2 below to check the balance between covariates for units from two treatment arms. The R-package `RItools` (function `plot.xbal`, Hansen and Bowers 2008) can be used to draw Love plots.

Another idea that can be borrowed directly from the matched sampling is checking the balance in propensity scores (Rosenbaum and Rubin 1983, 1985). Here we define a *propensity score* as $p = P(d = 1 \mid \boldsymbol{x}; \boldsymbol{\phi})$, a probability to be missing. After fitting the model for $d \mid \boldsymbol{x}, \boldsymbol{\phi}$, one can examine the overlap between empirical distributions of propensity scores for respondents and nonrespondents as well as test for differences in the distributions of $p|d = 1$ and $p|d = 0$ (possibly, on a logit scale) analytically.

Displaying multivariate data on a graph is especially challenging, and one plot type that does it effectively is parallel coordinate plot. It represents all variables (usually scaled to fall in $[0, 1]$ interval) by parallel vertical bars, and observations corresponding to each unit are connected by lines. These plots can be produced using the R-package `VIM` (Templ and Filzmoser 2008), which is devoted solely to creating various visualizations of missing values in a dataset to help explore their patterns. Figure 1.1 shows an example or the parallel coordinates plot, borrowed from (Templ and Filzmoser 2008).

The next two chapters describe a new systematic way to perform sensitivity analyses in studies with missing data by incorporating graphical displays.

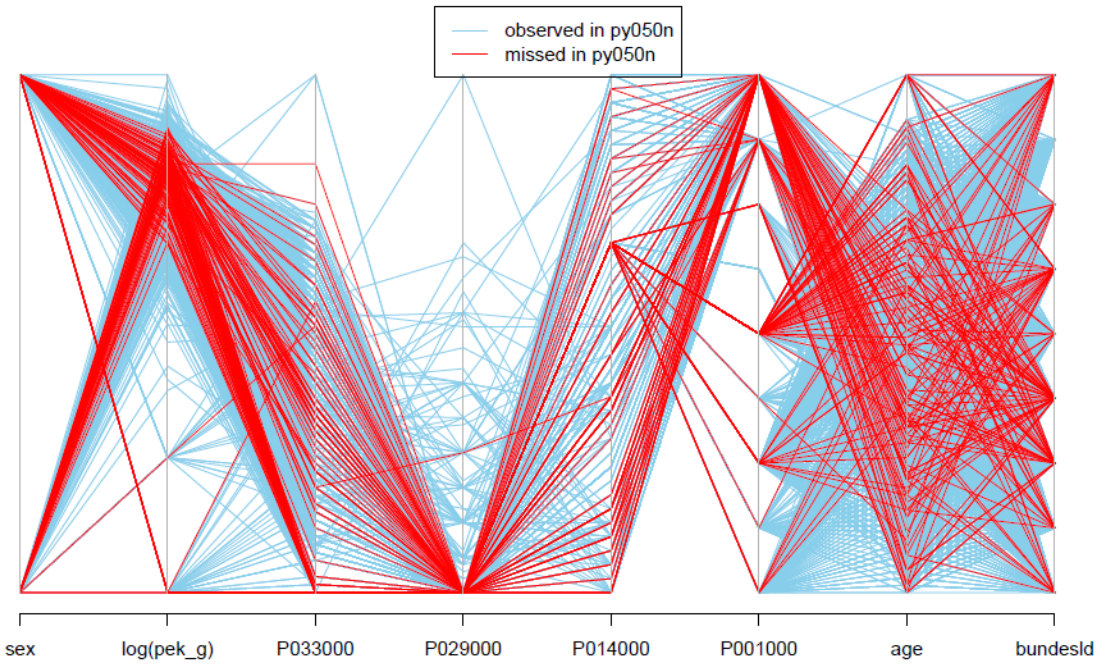


Figure 1.1: An example of the parallel coordinates plot taken from Templ and Filzmoser (2008). Here, the color indicates units with missing values in the variable *py050n*. We can notice that units with missing *py050n* have high portion of small values in the variable *P033000*, and $P029000 = 0$ for all of them. Also, some categories of *P001000* and *bundesld* do not have any units that are missing *py050n*, and, for the variable *pek_g*, nonrespondents fall only in a certain range.

Chapter 2

Sensitivity Analysis for Partially Missing Binary Outcomes in a Clinical Trial with Two Arms

2.1 Introduction

Various methods of handling data with missing values have been proposed in the literature. Each one of them requires assumptions about the missingness mechanism, implicit or explicit, and full appreciation was not given to the importance of these assumptions until the pivotal work of D. Rubin in the 1970s. As described in Section 1.1, Rubin (1976) proposed to treat missingness indicators as random variables, and, since then, three missingness mechanisms were defined, MCAR, MAR, and MNAR.

Here we focus on a special but very common case when the outcome data is partially missing and a set of fully-observed predictors that explain the missingness

and the outcomes is available. Let $\boldsymbol{Y} = (y_1, \ldots, y_N)'$, where y_i denotes a value of a univariate outcome of interest for unit i, and let $\boldsymbol{D} = (d_1, \ldots, d_N)'$ be the missingness indicator, such that $d_i = 0$ for units that are missing y_i and $d_i = 1$ for units with observed y_i. Let $\boldsymbol{X} = (x_{i,j})$ be a set of predictors that consists of three nonoverlapping subsets: predictors $\boldsymbol{X}_Y$ of the response $\boldsymbol{Y}$ only, predictors $\boldsymbol{X}_D$ of the missingness indicator $\boldsymbol{D}$ only, and common predictors $\boldsymbol{X}_{YD}$ for $\boldsymbol{Y}$ and $\boldsymbol{D}$, such that $\boldsymbol{X}_Y$, $\boldsymbol{X}_D$, and $\boldsymbol{X}_{YD}$ do not overlap. The triplet $(\boldsymbol{x}_i, y_i, d_i)$ is assumed to be independent and exchangeable across units, so we drop the index i to keep the notation in this section uncluttered.

Let the probability distribution of the outcomes for each unit be

$$f(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = f(y \mid \boldsymbol{x}_Y, \boldsymbol{x}_{YD}; \boldsymbol{\theta})$$

and the probability distribution of the missingness indicator be

$$f(d \mid \boldsymbol{x}, y, \boldsymbol{\phi}) = f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, y; \boldsymbol{\phi}),$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are vector-parameters governing the corresponding distributions. Then, each missingness mechanism defined in Section 1.1 implies that the following holds for every unit:

- MCAR: $f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, y; \boldsymbol{\phi}) = f(d \mid \boldsymbol{\phi})$ for each $\boldsymbol{\phi}$ and for all $\boldsymbol{x}$ and y. In other words, $\boldsymbol{X}_D$ and $\boldsymbol{X}_{YD}$ are empty sets and the missingness is independent of the response y itself.

- MAR: $f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, y; \boldsymbol{\phi}) = f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}; \boldsymbol{\phi})$ for the observed d, $\boldsymbol{x}$, and y, and

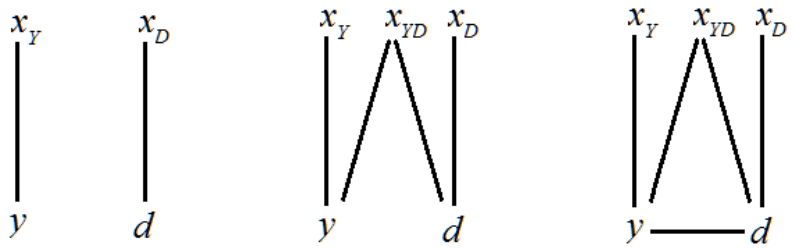
for each $\boldsymbol{\phi}$.

- MNAR: $f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, y; \boldsymbol{\phi}) \neq f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}; \boldsymbol{\phi})$. Note that MNAR can imply that there is an unobserved variable u that is associated both with the response and with the missingness indicator, such that $f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, u, y; \boldsymbol{\phi}) = f(d \mid \boldsymbol{x}_D, \boldsymbol{x}_{YD}, u; \boldsymbol{\phi})$, but, because we failed to measure u, the model for the missingness mechanism requires conditioning on the response y itself.

Definitions of all three missingness mechanisms do not assume anything about the distribution of the outcome y, so that it does not even have to be a random variable (Rubin 1976). However, here we specifically focus on the situation where the distribution of the outcome is modeled using covariates $\boldsymbol{X}$.

Figure 2.1 displays graphically the three mechanisms described above. The top row shows available predictors, and the bottom row shows outcomes. Conditional dependencies are represented by lines, while the absence of a line indicates conditional independence between the corresponding variables. Here, the dependency between variables is not limited to a linear model, as in Cox and Wermuth (1993), nor does a line suggest any causal relationship, as in Pearl (2009).

Many studies with missing data either use complete-case analysis (i.e., discard units with partially missing data), which is generally invalid, except in very special cases of MCAR mechanisms, or choose to analyze the data under the MAR assumption. The latter is usually a more sound approach than the former, especially when the MCAR assumption is contradicted by the observed data. At the same time, the MAR assumption allows us to avoid specifying a model for missingness mechanism for Bayesian or direct-likelihood inferences, assuming $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are distinct (see Section



(a) MCAR (b) MAR (c) MNAR

Figure 2.1: Illustration of the types of missingness mechanisms introduced in Section 1.1 for a special case with univariate outcome and no missingness in covariates. Panel (a) shows that, under MCAR, $\boldsymbol{x}_{YD}$ is empty and d is not related to y. Panel (b) indicates that, for MAR, d is allowed to be associated with y through the mutual predictors $\boldsymbol{x}_{YD}$. As evident from the diagrams, MCAR assumption is a special case of MAR . Finally, panel (c) shows that MNAR includes all cases that are not MAR.

1.1). However, although the MCAR assumption may be tested empirically (see Section 1.3.1, Rubin 1976; Little 1988b), the MAR assumption is generally unassessable. Therefore, a thorough sensitivity check is necessary to assess the influence of various assumptions about the missingness mechanism on study conclusions.

Here, focusing on binary outcomes, we describe a set of convenient displays that reveal the effects of all possible combinations of the values of missing data in treatment and control groups on various quantities of interest, typically, on p-values and point estimates. The displays are based on the idea of "tipping-point" analysis, first introduced in Yan et al. (2009), but anticipated in Matts et al. (1997), Hollis (2002), and Weatherall et al. (2009), as a method of assessing the impact of missing data on study's conclusions about some quantity of interest.

Tipping points of a study are defined as particular combinations of missing data values that would change the study's conclusions. Yan et al. (2009) presented a simple way to display these combinations for studies with two arms and a binary

outcome. We enhance this initial idea by adding more details onto the display. In particular, we allow for smooth changes in quantities of interest, add the output from multiple missingness models, including MNAR, and, when available, mark historical estimates. We show how the display can help to systematize the sensitivity analyses and to demonstrate the results across different alternative models. The proposed displays enable practitioners to identify how close alternative assumptions about the missingness mechanism come to altering the study's conclusions and, thereby, to assess the strength of the study's evidence.

The rest of the chapter is organized as follows. Section 2.2 lays out the basics of the sensitivity analysis and the motivation for the proposed technique. Section 2.3 provides a detailed description of enhanced tipping-point displays for a binary outcome. It also includes a simulated example that demonstrates the technique and a real-data example of the recent use of the enhanced displays in a medical device clinical trial. We conclude with a discussion (Section 2.4).

2.2 Sensitivity Analyses for Studies with Missing Data

In every empirical study plagued with missing data, researchers face a tough decision about the method of handling it. The choice of the method should be justified by stating and discussing the required assumptions and, possibly, applying alternative methods to assess the extent to which the study conclusions depend on the assumptions used. The latter constitutes the essence of a *sensitivity analysis* for studies with

missing data, which is especially necessary if the assumptions about the missingness mechanism used in the study are unassessable, which is typical.

A sensitivity analysis consists of several steps:

- Formulating conclusions under working assumptions;
- Identifying a set of plausible alternative assumptions;
- Studying the variation in the statistical output and conclusions under these alternative settings.

Because many methods for handling missing assume a MAR mechanism, the last two steps imply weakening this assumption. However, the apparent complexity of MNAR models appears to be the primary reason why the majority of empirical research chooses to omit any sensitivity analysis altogether. Yet, in some cases, omitting it is not an acceptable option, especially when it comes to important decisions like approving a drug or a medical device, or implementing a new public policy. For example, NAS report on methods of handling missing data (NRC-Panel 2010, p. 5) made the following recommendation: "*Recommendation 15: Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.*" Other guidelines issued lately (Burzykowski et al. 2010; CHMP 2010) also stressed the need to perform sensitivity analyses that assess the impact of missing data on reported inferences and conclusions.

In spite of the rising demand, there is clearly a shortage of practical recommendations as to how one should perform sensitivity analyses (Lee 2007; NRC-Panel 2010).

As pointed out in NRC-Panel (2010, p. 83), "*Unlike the well-developed literature on drawing inferences from incomplete data, the literature on the assessment of sensitivity to various assumptions is relatively new. Because it is an active area of research, it is more difficult to identify a clear consensus about how sensitivity analyses should be conducted.*" We address this issue below and demonstrate a process of exploring MAR and MNAR models for studies with missing values in binary outcomes using enhanced tipping-point displays.

2.3 Enhanced Tipping-Point Displays for Studies with a Binary Outcome

Tipping-point (TP) analysis was first proposed in Yan et al. (2009) to aid clinical reviewers in judging the impact of missing data in the outcome on the estimation of a treatment effect. Yan et al. (2009) constructed displays to help illustrate "tipping points" of a study, i.e., the combination of possible values of missing outcomes that would reverse the conclusion about the statistical significance of the treatment effect. These displays were further discussed in Campbell et al. (2011) as a convenient tool to reveal the results of sensitivity analysis to various deviations from assumptions made about the missing data mechanism.

Suppose that a study is conducted to estimate the effect of a vaccine (or a *treatment*) on a subsequent occurrence of a disease. A total of N study subjects are divided into treatment group or control group, and a $(2 \times N)$ set of predictors $\boldsymbol{X}$, along with a vector of treatment indicators $\boldsymbol{T} = (t_1, \ldots, t_N)'$, are completely observed

for all subjects. A vector of outcomes $\boldsymbol{Y} = (y_1, \ldots, y_N)'$ indicates whether each subject developed the disease ("success") or not ("failure") and some subjects are missing the outcome, as indicated by the vector of missingness indicators $\boldsymbol{D} = (d_1, \ldots, d_N)'$. Vector $\boldsymbol{Y}$ has four parts that correspond to observed and missing outcomes among treatment and control subjects, i.e., $\boldsymbol{Y}^T_{obs}$, $\boldsymbol{Y}^C_{obs}$, $\boldsymbol{Y}^T_{mis}$, and $\boldsymbol{Y}^C_{mis}$, such that

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}^T \\ \boldsymbol{Y}^C \end{pmatrix}, \boldsymbol{Y}^T = \begin{pmatrix} \boldsymbol{Y}^T_{obs} \\ \boldsymbol{Y}^T_{mis} \end{pmatrix}, \boldsymbol{Y}^C = \begin{pmatrix} \boldsymbol{Y}^C_{obs} \\ \boldsymbol{Y}^C_{mis} \end{pmatrix}.$$

Let $\tau = E(y_i \mid t_i = 1, \boldsymbol{\theta}) - E(y_i \mid t_i = 0, \boldsymbol{\theta})$ be a marginal average treatment effect, identical for all subjects $i = 1, \ldots, N$. If the treatment is properly randomized between the subjects, an unbiased estimator of τ is

$$\hat{\tau} = \sum_{i: y_i \in \boldsymbol{Y}^T} y_i / N^T - \sum_{i: y_i \in \boldsymbol{Y}^C} y_i / N^C = \bar{y}^T - \bar{y}^C, \tag{2.1}$$

where N^T and N^C are the sample sizes for treatment group and control group respectively.

For a binary outcome $\boldsymbol{Y}$, an intuitive summary of missing values is the number of successes among subjects with missing outcomes, considered separately for treatment group and control group,

$$g(\boldsymbol{Y}^T_{mis}) = \sum_{i: y_i \in \boldsymbol{Y}^T_{mis}} y_i = N^T_{mis} \bar{y}^T_{mis}, \quad g(\boldsymbol{Y}^C_{mis}) = \sum_{i: y_i \in \boldsymbol{Y}^C_{mis}} y_i = N^C_{mis} \bar{y}^C_{mis},$$

where $N^T = N^T_{mis} + N^T_{obs}$ and $N^C = N^C_{mis} + N^C_{obs}$. Moreover, these summaries are

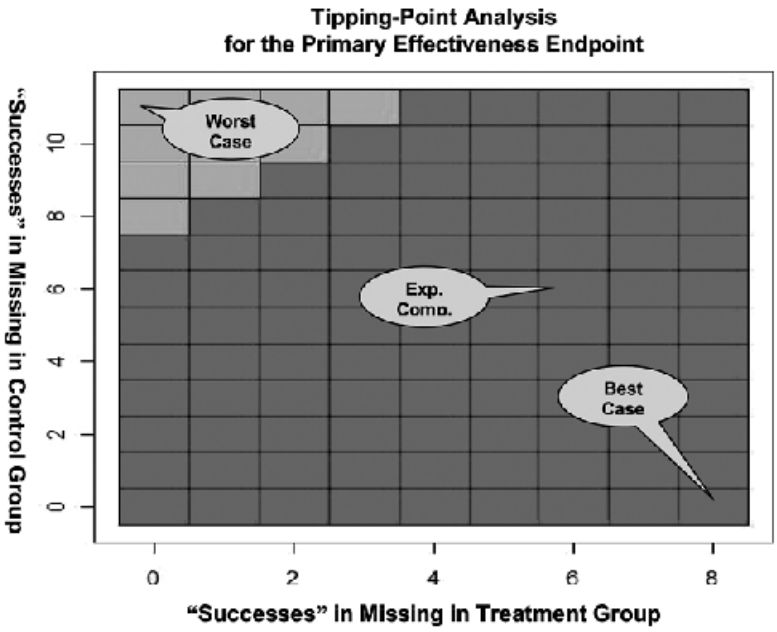


Figure 2.2: This illustration is taken from Campbell et al. (2011) to demonstrate the idea proposed in Yan et al. (2009). The horizontal and vertical axes indicate the number of successes that can potentially be observed among nonrespondents in the treatment group and the control group. Each combination is marked as either "altering the study's conclusion" (lighter squares) or "keeping the study's conclusion unchanged" (darker squares). The staircase region indicates the tipping points of the study.

justified by the fact that, for N i.i.d binary variables with probability of success p,

$$y_1, \ldots, y_N \mid p \sim \text{Bern}(p),$$

a minimum sufficient statistic (MSS) for estimating p is $\sum_{i=1}^{N} y_i$. Therefore, with respect to this model, no information is lost by collapsing missing outcomes into one summary in each group. Therefore, $g(\boldsymbol{Y}_{mis}^{T})$ and $g(\boldsymbol{Y}_{mis}^{C})$ can be represented by the two axes of the enhanced TP display.

Figure 2.2 from Campbell et al. (2011) illustrates the initial idea of a tipping-point

display described in Yan et al. (2009) for a binary outcome, where it results in a matrix of all possible combinations of the number of successes among nonrespondents in the treatment group and in the control group. Each combination is categorized based on whether the corresponding missing pattern changes, or "tips", the conclusion about the estimated effect's statistical significance. The staircase region marks the tipping points of the study, i.e., the combinations of the number of successes among nonrespondents in the treatment group (horizontal axes) and in the control group (vertical axes) that alter the conclusion about the statistical significance of the treatment effect, based on a chosen hypothesis test and a significance level. One fundamental issue with this basic depiction is that the display has no information about the likelihood of each individual combination. Therefore, unless we discover that none of possible missing data patterns change the study's conclusion, we cannot utilize these displays to their fullest potential.

We use the initial idea of illustrating tipping points to propose a visual approach to performing sensitivity analysis. It is done by introducing the following enhancements to the displays:

- A colored heat-map that illustrates the *gradual* change of the quantity of interest, e.g., the p-value from a hypothesis test used in the study. Moreover, it can also represent the estimated treatment effects, $\hat{\tau}$, the lower or upper bounds of confidence interval, or any other quantity that depends on a particular combination of the number of successes among nonrespondents in the treatment group and in the control group.

- Ticks, which represent historical estimates of the number of successes in each

group, if such are available. For example, if axes represent the number of adverse events among treated and among control subjects, the ticks could indicate the numbers that correspond to the rates observed in previous studies for patients with similar demographics and medical condition.

- The results from the current modeling procedure, e.g., the posterior draws of $\boldsymbol{Y}_{mis}$ under the chosen model $f(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\phi})$.

- Most important, the posterior draws of $\boldsymbol{Y}_{mis}$ obtained under models with alternative assumptions utilized for the sensitivity analysis. We elaborate on the last two enhancements in the following sections.

The merit of adding ticks that correspond to historical and observed values is especially apparent because the practitioner may compare them with the values obtained under the primary and alternative models and, based on that, judge the sensibility of underlying assumptions.

As already mentioned, there are several quantities that may be of interest to a practitioner and could be represented by a heat-map on a TP display. First, it can represent the estimate of τ, as it varies depending on the number of successes among missing outcomes. The relationship may be expressed as follows:

$$\begin{aligned} \hat{\tau} &= \frac{\bar{y}^T_{obs} N^T_{obs} + \bar{y}^T_{mis} N^T_{mis}}{N^T} - \frac{\bar{y}^C_{obs} N^C_{obs} + \bar{y}^C_{mis} N^C_{mis}}{N^C} \\ &= \frac{\bar{y}^T_{obs} N^T_{obs} + g(\boldsymbol{Y}^T_{mis})}{N^T} - \frac{\bar{y}^C_{obs} N^C_{obs} + g(\boldsymbol{Y}^C_{mis})}{N^C}. \end{aligned} \tag{2.2}$$

Another quantity of interest is the p-value that corresponds to a test of the estimated treatment effect $\hat{\tau}$. Next, we illustrate the use of enhanced TP (or ETP) displays on

a simulated example with a binary outcome and several fully-observed predictors.

2.3.1 Simulated Example with a Binary Outcome

In order to illustrate the use of ETP displays with a binary outcome, we generated data for $N = 100$ subjects with two predictors, representing sex, $\boldsymbol{Female} = (female_1, \ldots, female_N)'$, and age in years, $\boldsymbol{Age} = (age_1, \ldots, age_N)'$, a treatment indicator $\boldsymbol{T} = (t_1, \ldots, t_N)'$, and a partially missing outcome $\boldsymbol{Y} = (y_1, \ldots, y_N)'$ (adverse event occurrence). Predictor ***Female*** was simulated from Bern(0.5), and predictor ***Age*** was simulated uniformly between 18 and 55 (rounding to the nearest integer).

The following models were used to generate the outcomes and the missingness,

$$\begin{aligned} \text{logit}(p_i) =& 2t_i - 0.001age_i - 0.1female_i \\ & - 0.05female_i \cdot age_i \cdot I(age_i > 35) \\ & - 0.001female_i \cdot age_i^2 \cdot I(age_i > 35), \end{aligned} \tag{2.3a}$$

$$y_i \mid p_i \sim \text{Binom}(p_i), \tag{2.3b}$$

$$\text{logit}(e_i) = 3 - 0.1age_i - 0.5female_i + 0.5y_i, \tag{2.3c}$$

$$d_i \mid e_i \sim \text{Binom}(e_i),\ i = 1, \ldots, N, \tag{2.3d}$$

where $I(\cdot)$ is an indicator function. According to the notation introduced in Section 2.1, here $\boldsymbol{X}_{YD} = (\boldsymbol{T}, \boldsymbol{Age}, \boldsymbol{Female})$, while $\boldsymbol{X}_Y$ and $\boldsymbol{X}_D$ are empty. As evident from (2.3c), the missingness mechanism is MNAR. The model for p_i (2.3a), the probability of success for subject i, indicates that, although the treatment effect is positive, the success rates decline steeply for females over 35. The rapid increase in the risk of

adverse events after reaching a certain age is not an uncommon phenomenon, e.g., the risk of heart disease increases for men after the age of 45 and for women after the age of 55, the risk of having fertility issues (miscarriage, birth defects, etc.) increase sharply for women over 35.

In the simulated data, out of 100 subjects, $N^T = 40$ were randomly assigned to the treatment group and $N^C = 60$ to the control group, with $N^T_{mis} = 15$ and $N^C_{mis} = 21$ subjects missing the outcome in each group, respectively. Figure 2.3 shows the heat-map of $\hat{\tau}$ for the generated data set, calculated according to (2.2). If we perform the hypothesis test for the difference in proportions of successes between treatment group and control group, the results may also be shown on the ETP display. Figure 2.4 shows the heat-map of p-values and outlines the region that corresponds to a significant treatment effect based on the significance level of 0.05. Hence, the outer contour of the region indicates the tipping points of the study, i.e., the number of successes among missing outcome values in treatment group and control group that would change the conclusion of the study e.g., {1,0},{2,0},{2,1} etc. Undoubtedly, the best possible scenario for a researcher would be when the display shows no tipping points, i.e., when all combinations of missing outcomes lead to the same conclusion of the study. If it is not the case, as in our simulated example, then performing sensitivity analysis can be critical, and ETP displays can be used to guide it.

Next, we illustrate the results of three analyses performed on the simulated data. The first analysis assumes MCAR model and multiply imputes the missing outcomes based on the rates of adverse events observed among respondents, without taking into account available predictors. The last two analyses assume a MAR

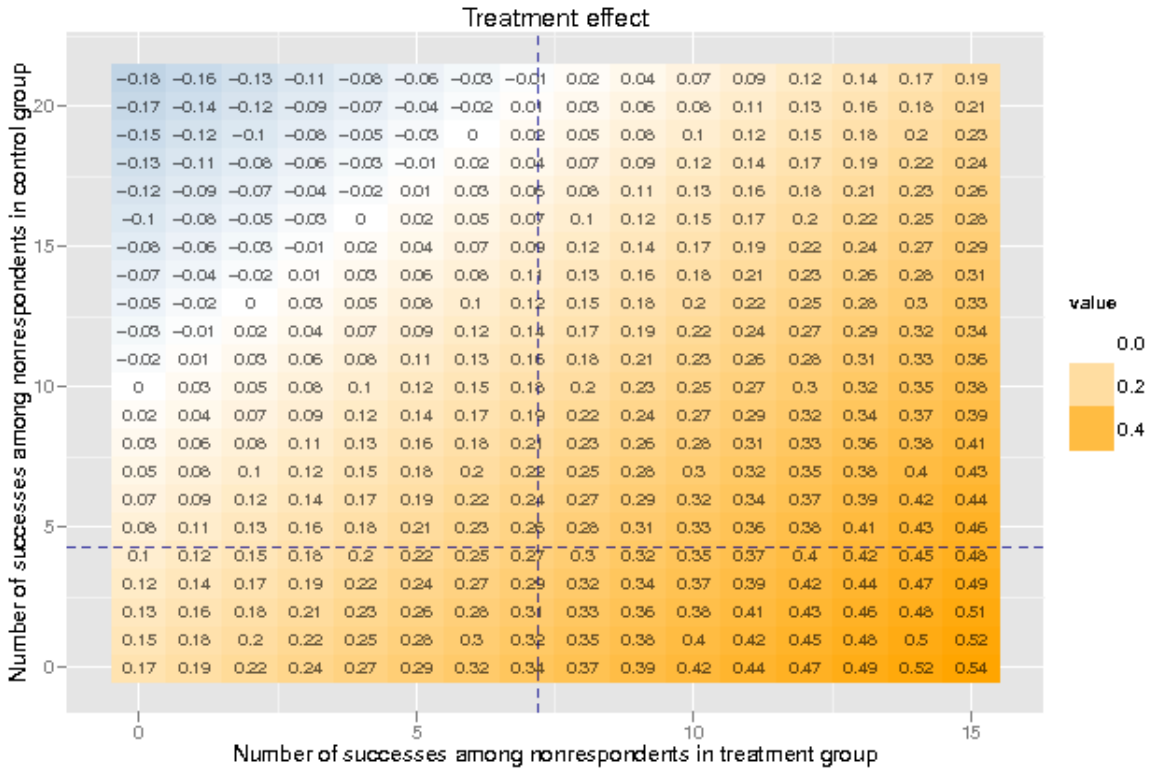


Figure 2.3: ETP display for the simulated binary outcome $\boldsymbol{Y}$, showing the estimated treatment effects using a heat-map. Axes represent the number of successes that could be observed among nonrespondents in the treatment group and in the control group. Each combination corresponds to a value of the estimated treatment effect $\hat{\tau}$ according to (2.2). Its magnitude and sign are represented using a color palette that changes from dark blue (large negative value) to dark orange (large positive values), with white representing zero estimated effect. Note that displaying each individual value is optional (and, in fact, largely redundant), so we omit it in further displays. The axes indicate that there were 15 missing outcomes among treated subjects and 21 among control subjects. Vertical and horizontal dashed lines (in blue) correspond to observed success rate among treated and control subjects, 0.48 and 0.21.

mechanism, and multiply impute missing values from their approximate posterior predictive distributions, obtained using MICE algorithm. The second analysis uses a naïve linear model for the log-odds of success to impute missing responses, i.e, $\text{logit}(p_i \mid t_i, age_i, female_i; \boldsymbol{\theta}) = \theta_0 + \theta_1 t_i + \theta_2 age_i + \theta_3 female_i$. The third analysis

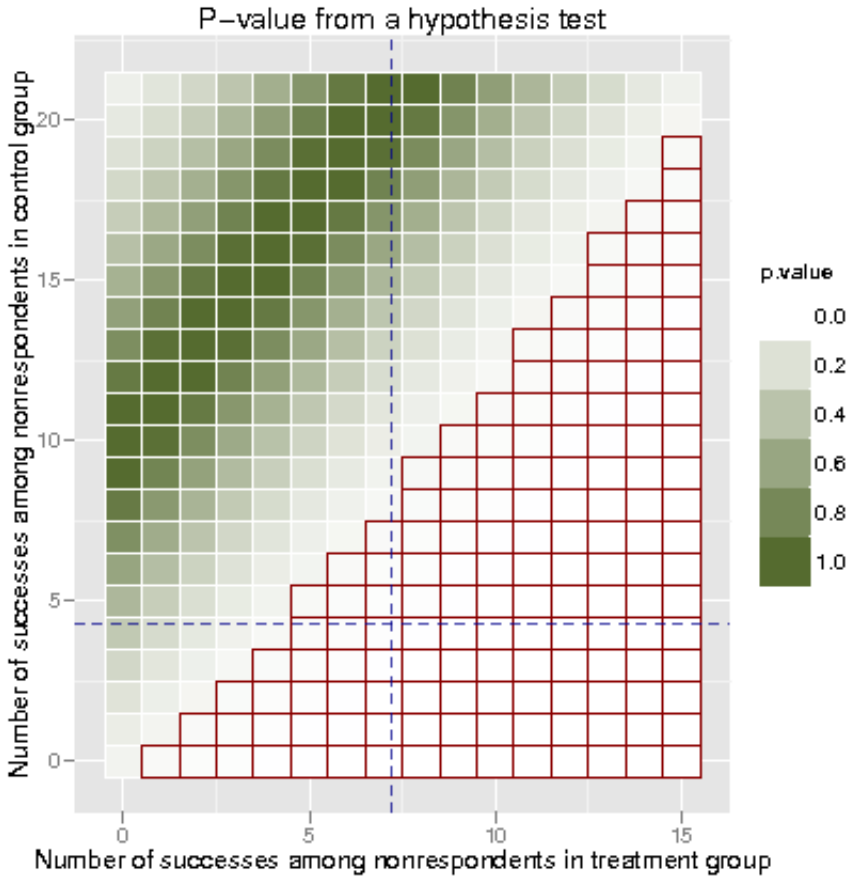


Figure 2.4: ETP display for the simulated binary outcome $\boldsymbol{Y}$, showing the p-values from a chosen hypothesis test (i.e., test of the difference in proportions of successes in treated and control groups). The heat-map represents p-values obtained from the test conducted for each combination of the number of successes among treated and among control subjects. The red grid highlights combinations that result in a significant treatment effect at the 0.05 significance level, with a stair-case region indicating the tipping points of the study.

includes all the relevant interactions, as specified in (2.3a) and, therefore, is more accurate. Note that the actual details of the imputation procedure are not essential, as long as a the procedure is proper and it uses plausible assumptions about the missingness mechanism.

Table 2.1 gives the estimate of the treatment effect for the full data set. It also

Table 2.1: Treatment effect on the outcome $\boldsymbol{Y}$, estimated for the full dataset as well as for the observed dataset, with missing values multiply imputed using three models. For the naïve and the complete models we assume MAR missingness. The results from 100 MIs are combined for each model using Rubin's rule.

Analysis	Estimated difference	95% Interval
Full data	0.27	(0.09, 0.46)
MCAR	0.27	(0.05, 0.48)
Naïve model	0.24	(-0.04, 0.53)
Complete model	0.31	(0.05, 0.57)

gives the estimates and corresponding 95% credible intervals obtained from 100 MIs generated for each of the three analyses and combined using the Rubin's rule (Rubin 1987; Barnard and Rubin 1999). Figures 2.5 and 2.6 show the results of the MI procedures, demonstrating different ways that the joint posterior distribution of the missing values can be summarized[1]. Brown, blue, and red rectangles are drawn by connecting minimum and maximum values among the imputations in each group under the naïve, complete, and MCAR models, respectively. In Figure 2.6, the (jittered) points indicate actual imputed values for each model. The corresponding contours encircle 95% of points for each model, obtained by excluding 5% of points that have the largest Mahalanobis distance from the sample mean. These contours approximates the 95% posterior region of the joint distribution of successes among nonrespondents in the treatment group and the control group.

We also added several vertical and horizontal ticks, showing counts that correspond to hypothetical historical data. For example, if rates of success for subjects with similar demographics were observed to be 0.35 and 0.60 in previous studies of similar treatments, for our example they would correspond to having 2 and 12

[1]The R-procedure that constructs ETP displays for generated MIs can be downloaded from http://sites.google.com/site/vliublinska/research.

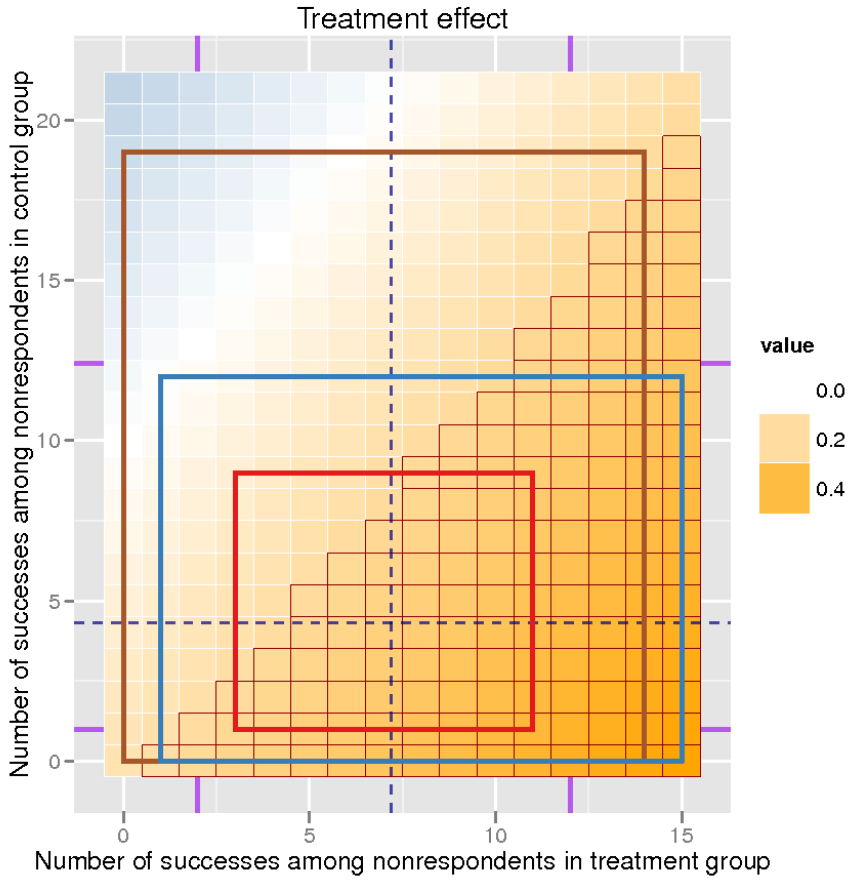


Figure 2.5: ETP display showing results from three MI procedures for the simulated binary outcome $\boldsymbol{Y}$. As before, the red grid highlights combinations that correspond to a significant treatment effect based on a hypothesis test for the difference between two proportions, using 0.05 significance level. In this simple version of the ETP display, the rectangles indicate minimum and maximum values among 100 imputed numbers of successes for nonrespondents in the treatment group and the control group under the naïve (brown), the complete (blue), and the MCAR (red) models. Also, the display shows two vertical and two horizontal ticks (in purple), representing counts that correspond to success rates $\{0.35, 0.60\}$ for the treated, and $\{0.15, 0.34\}$ for the controls, serving to illustrate the use of data possibly available from previous studies. This version of the ETP display (with heat-map showing p-values instead of treatment effects) is used in the real-data example in Section 2.3.2.

successes among nonrespondents in the treatment group, respectively.

Figures 2.5 and 2.6 reveal a difference between counts imputed using the three

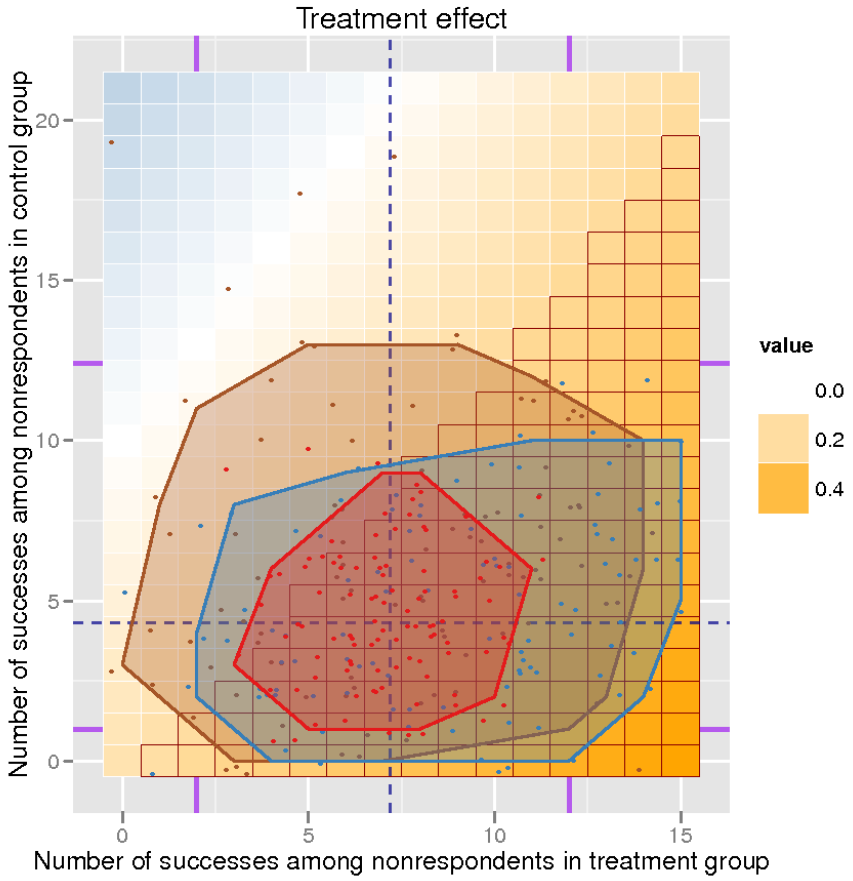


Figure 2.6: ETP display, similar to the one shown in Figure 2.5, but more detailed. The jittered points indicate the number of imputed successes for nonrespondents in the treatment group and the control group under the naïve (brown), the complete (blue), and the MCAR (red) models. Brown, blue, and red contours contain 95% of the imputations, while 5% of points with the largest Mahalanobis distance from the sample average are excluded. The contours approximate the 95% posterior region of the joint distribution of the number of successes among nonrespondents in the treated group and the control group. The results obtained from the models are somewhat different, indicating that both naïve and MCAR models may not be accurate.

models. In addition, Table 2.1 shows that the three models produce conflicting conclusions regarding the significance of the effect, with the naïve one indicating that there is no significant treatment effect. If additional predictors in the complete model were not relevant, we would expect similar results to be produced under both models.

Next we describe how a systematic sensitivity analysis was performed on a real data from a medical device clinical trial with multiple binary outcomes and substantial missingness, and how ETP displays were utilized to summarize it.

2.3.2 Real-data Example

So far we focused on the situation with missing values confined to a single outcome. However, the example that we present next involves a more complex problem and demonstrates how the TP analysis can be extended to the situation with missingness in more then one outcome. The data set that we use comes from a clinical trial conducted in 2008-2009 in Germany. The objective of the study was to compare the efficacy and safety of a new device for kyphoplasty, a novel treatment of vertebral compression fractures, which are the most common complications of osteoporosis, to the efficacy and safety of a traditional procedure, i.e., vertebroplasty. Both procedures involve the injection of bone cement into fractured vertebrae, with the goal to relieve pain caused by their compression and to prevent further damage.

A randomized prospective open-label study took place in four health centers across Germany. The inclusion criteria for patients required, among other things, to have up to three vertebral compression fractures in a specific region of their spines, to be at least 50 years old, and to have pain levels above a certain threshold. A total of 84 subjects were evaluated, qualified, consented and randomized to one of the two procedures, yielding 56 subjects assigned to the kyphoplasty ("treatment" group) and 28 to the vertebroplasty ("control" group).

A primary endpoint of the study was the number of cement leaks into the spinal

canal, a potentially extremely serious complication that may lead to paraplegia. This endpoint, as well as the pain score, were collected 24 hours after the surgery, while the patients were still in the hospital. Both variables did not have any missing data, therefore, we will not focus on them in this section. However, a randomization-based analysis of these endpoints were highly supportive of the superiority of the kyphoplasty procedure, performed using the new device.

The study also had several secondary endpoints, including the occurrence of various adverse events within 3 months and between 3 and 12 months after the procedure, that assessed the relative safety of the new device. The following six types of adverse events were studied:

- adjacent level vertebral fracture (symptomatic and asymptomatic),
- distant level vertebral fracture (symptomatic and asymptomatic),
- retreatment (including refracture),
- death (12-month observations include deaths within 3 months).

In addition, subjects' pain levels (0 through 10) and disability scores (0 through 100, assigned based on a completed questionnaire) were recorded during the 3- and 12-months follow-up appointments. Table 2.2 summarizes all secondary endpoints collected in the study. In addition, a set of baseline measurements was collected for each randomized patient, including:

- the number of vertebral compression fractures that required treatment (1, 2 or 3),

Table 2.2: Secondary endpoints collected in the study, indicated by "+".

Secondary Endpoint	Time after surgery		
	At 24h	1 day to 3 months	3 to 12 months
Occurrence of each of the six adverse events		+	+
Pain level (0-10)	+	+	+
Disability score (0-100)		+	+

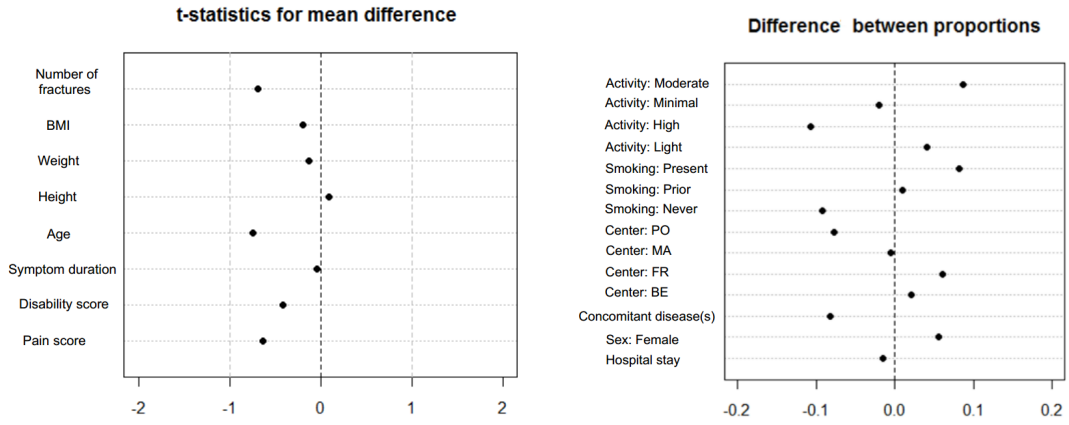
- demographic and health data (age, sex, height, weight, BMI, physical activity level, smoking status),
- baseline pain and disability scores, duration of symptoms, health center of stay, presence of concomitant disease(s).

A considerable fraction of subjects were missing secondary endpoints. Table 2.3 reports percents of subjects in each group that had missing outcomes at each time-point. Also, the occurrence of adverse events was rare, with the range of observed rates between 0% and 2.6%, with the exception of deaths that were reported at 10.4% rate during the 12-months follow-up; the patients' age range at the baseline was 50 to 93, therefore such a high death rate was not surprising. However, death is considered to be unrelated to the treatment assigned. In addition, a few subjects had missingness in one or more of the baseline covariates. In summary, the study had several major issues that complicated the analysis and required thorough attention: considerable fraction of non-monotone missing data in secondary outcomes that were rarely occurring events, some missingness in covariates and, moreover, small sample sizes in the treatment and the control groups. Therefore, regardless of what assumptions about the missingness we used for the initial analysis, it was essential to perform a thorough sensitivity check to these assumption, and ETP displays were utilized for

Table 2.3: Percent of subjects missing all secondary endpoints at each time-point.

Treatment group	Follow-up time-point		
	3 months, %	12 months, %	3 & 12 months, %
Kyphoplasty ($N^T = 49$)†	24	43	18
Vertebroplasty ($N^C = 28$)	18	36	11

†7 subjects were excluded from the treatment group after randomization due to issues unrelated to the actual procedure.



(a) Continuous variables

(b) Discrete variables

Figure 2.7: Love plots to check the balance between the treatment group and the control group produced by the randomization.

this purpose.

We start with assessing the randomization procedure and making sure it produced an acceptable balance between the treatment group and the control group. Figure 2.7 contains two "Love plots", described in Section 1.3.2 (Ahmed et al. 2006), that show standardized differences between average values of baseline measurements, or between proportions for binary measurements, observed in each group. The two plots indicate an excellent balance across the two groups. We proceed with multiply imputing few missing values in baseline covariates. For that, we combine the two groups, as

justified by the randomization, but remove the outcome data. We assume MAR missingness in baseline measurements and apply the MICE algorithm to produce 100 complete data sets that will be utilized in subsequent analyses. Next, we describe the adopted assumptions about the missingness mechanism for the secondary endpoints, the procedure used for estimating the treatment effect, and the obtained results.

Questions of interest that concern secondary endpoints are whether the two treatments differ in the rates of adverse events as well as in the post-treatment pain levels and disability scores. As noted above, all secondary endpoints had large proportions of missingness. Therefore, in order to perform the analysis, we have to consider plausible assumptions about their missingness mechanism. For the initial analysis we assume the MAR mechanism and proceed to multiply impute the missing secondary outcomes using the MICE algorithm, taking into account available baseline covariates. For that, the outcome data collected post-operatively are split into treatment group and control group. Two analysts are assigned to perform multiple imputation procedure on each part separately; both are blinded to each other's outcome data. This is done to limit the opportunity to bias the results, e.g., systematically impute better values for subjects in the treatment group, as well as to allow different response functions for the effects of each of the two treatments.

The sparsity of rare adverse events requires a special method of conditional imputation because it is not feasible to model the occurrence of each of the twelve adverse events (six types observed at two time-points) individually. Instead, we use a hot-deck approach by adopting a file-concatenation matching method introduced in Rubin (1986), where each subject with missing secondary outcomes (i.e, a non-

respondent) is matched based on available characteristics to a donor from a pool of respondents, and the entire set of outcomes from the found donor is used to impute missing outcomes for that nonrespondent. In addition, post-treatment pain scores and disability indexes collected during the 3- and 12-months follow-up appointments cannot be modeled as continuous variables due to small sample sizes and irregular distributions of the observed values. Therefore, for the purpose of multiple imputation, we employ predictive mean matching (PMM, Rubin 1986; Little 1988a), another hot-deck-type method that fits a linear model to observed responses and uses it to match each nonrespondent with respondents by comparing their predicted responses.

In order to test whether or not the treatment group and control group showed similar results in secondary outcomes, we employ a one-sided Fisher randomization test. Table 2.4 reports results obtained from 100 complete data sets, combined using Rubin's rule, as described in Licht (2010). The results support the conclusion that there is essentially no evidence that kyphoplasty, performed using the new device, is worse than vertebroplasty in the rate of any adverse event, as well as in average post-treatment pain scores or disability indexes. Next, we subject these conclusions to a thorough sensitivity assessment.

The unassessable MAR assumption that underlies the imputation model for missing secondary endpoints raises concerns due to the large fraction of missingness. As noted above, the imputation methods were hot-deck, i.e., using observed outcomes from respondent donors. Hence, an implicit assumption of such methods is that each nonrespondent resembles one or more of the respondents. However, further analysis revealed that there was some nonoverlap in the values of baseline measurements

Table 2.4: One-sided p-values from a Fisher randomization test for null-hypotheses of no difference between the treatment group and the control group in the rate of each of the adverse events. A one-sided alternative hypothesis was used to make it possible to combine p-values from 100 complete data sets (see Licht 2010). Note that none of the p-values provide any evidence against the corresponding null-hypotheses.

Alternative Hypothesis	Treated subjects have fewer adverse events	
Adverse events	With 3 months	Between 3 and 12 months
Retreatment	1.00	1.00
Symptomatic Adjacent Fracture	0.30	1.00
Symptomatic Distant Fracture	0.99	0.27
Asymptomatic Adjacent Fracture	1.00	0.99
Asymptomatic Distant Fracture	1.00	0.48
Death	0.13	0.59
Any event before 3 months	0.29	0.32
Pain score	0.66	0.29
Disability index	0.26	0.19
Alternative Hypothesis	Treated subjects have more adverse events	
Adverse events	With 3 months	Between 3 and 12 months
Retreatment	0.39	0.99
Symptomatic Adjacent Fracture	0.89	0.46
Symptomatic Distant Fracture	0.38	0.99
Asymptomatic Adjacent Fracture	1.00	1.00
Asymptomatic Distant Fracture	1.00	0.90
Death	0.99	0.68
Any event before 3 months	0.83	0.80
Pain score	0.34	0.71
Disability index	0.75	0.82

between respondents and nonrespondents in the control group (see Section 1.3.2). Specifically,

- At 3-months follow-up:
  - All three male nonrespondents were older than the oldest male respondent (76, 77, 83 vs. 69 years old at the beginning of the study);
  - Two out of three male nonrespondents had lower BMI than the lowest observed BMI among respondents (21.5, 20 vs. 23.5);

  - One out of two female nonrespondents had prior smoking experience, and no female respondent had it;
  - One male nonrespondent had a longer hospital stay duration than all male respondents.

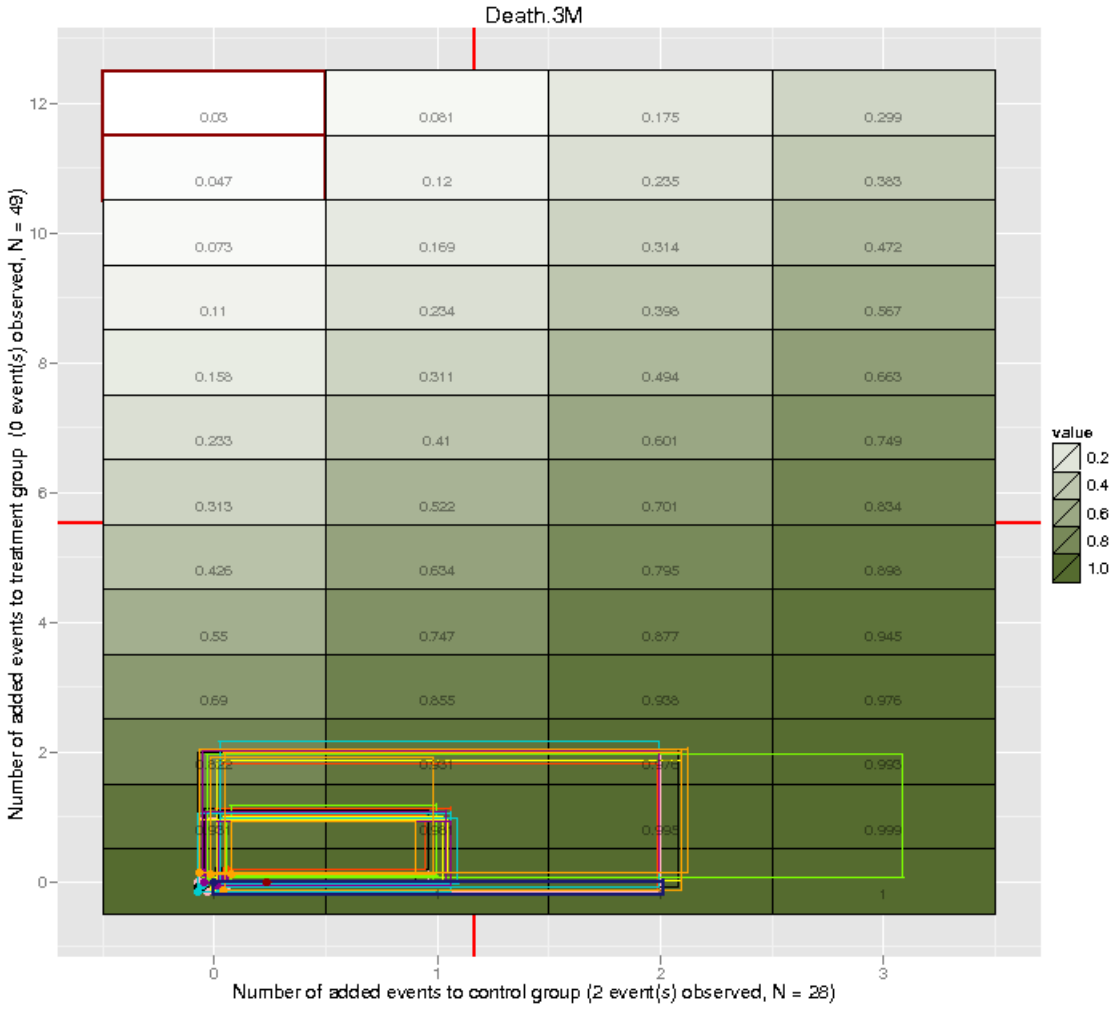
- At 12-months follow-up:
  - Two female nonrespondents were older than the oldest female respondent (88, 89 vs. 85);
  - One male nonrespondent was older than the oldest male respondent (83 vs. 77).

Note that the nonrespondents that did not resemble any respondents in the control group appeared to be in a poorer health than the respondents, e.g., older, with higher BMI etc. Consequently, by using responses from healthier subjects in the control group to impute missing outcomes for nonrespondents, the hot-deck imputation procedure produces results favoring the control group. Nevertheless, the detection of nonoverlap provided us with a direction for constructing MNAR models: identify specific characteristic of nonrespondents that are outside of the range observed among respondents and modify the odds of adverse events for subjects with these characteristics, taking the odds estimated under the MAR model as a baseline,

$$\text{logit}\{P(y_i = 1 \mid d_i = 1, t_i, \boldsymbol{x_i}, \boldsymbol{\theta})\} = \\ \text{logit}\{P(y_i = 1 \mid d_i = 0, t_i, \boldsymbol{x_i}, \boldsymbol{\theta})\} + \delta(t_i, \boldsymbol{x_i}),\ i = 1, \ldots, N.$$

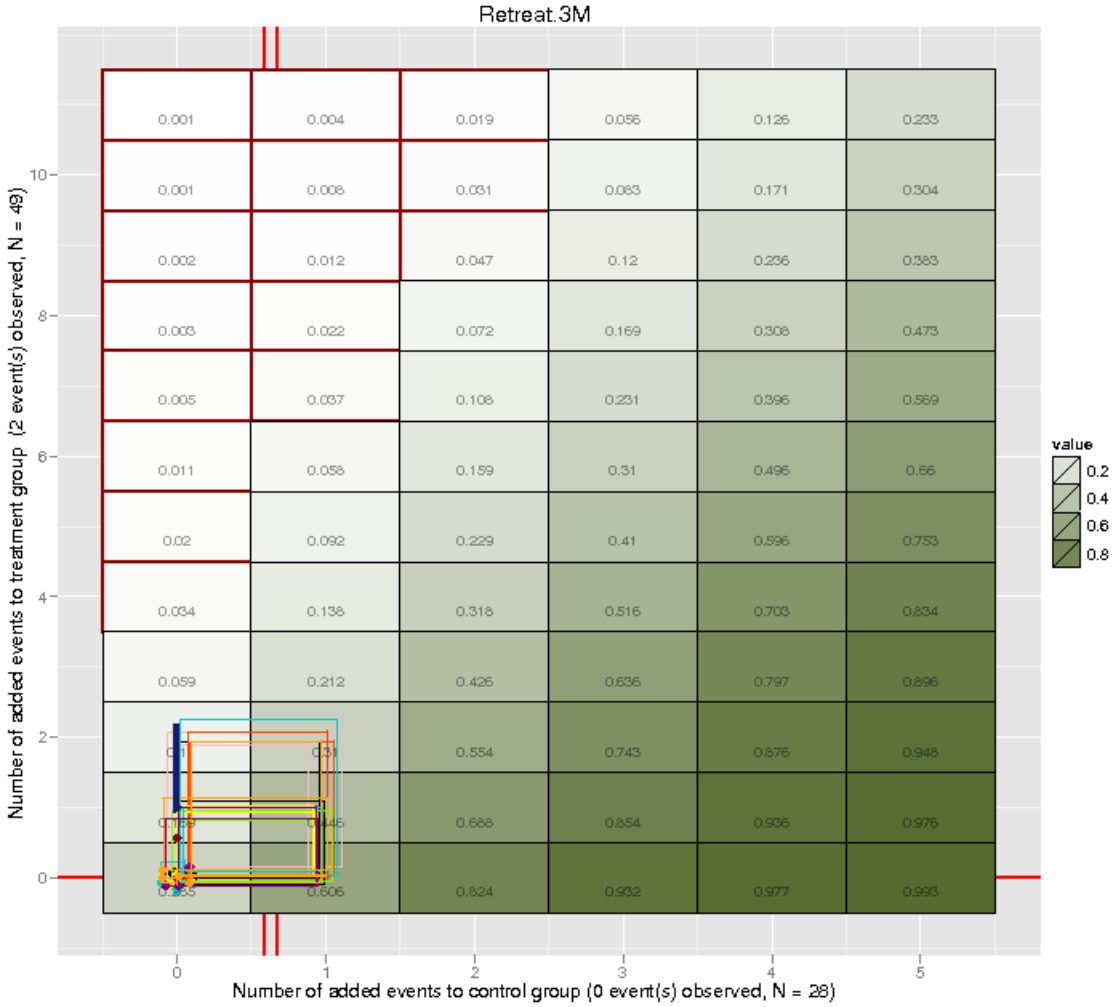
The following eight characteristics were selected for the purpose of the sensitivity analysis: males older than 69, males with BMI lower than 23.5, females with prior smoking experience, males with duration of hospital stay longer than 2 days, females older than 85, males older than 77, patients dead at 3 months, patients dead at 12 months. The odds of adverse events were imputed to be 50% higher ($\delta = \ln(1.5)$) or 50% lower ($\delta = \ln(0.5)$) than implied by the MAR model for the treatment group or the control group separately. A total of 32 alternative models were fitted (eight characteristics for two groups and two odds adjustments) and 100 MIs were produced for each of them. Similarly to the simulated example on Figure 2.5, Figure 2.8 shows the resulting ETP displays with rectangles indicating ranges of the number of adverse events imputed under the initial model with the MAR assumption (dark blue) as well as under each of the 32 alternative models. The heat-map represents p-values from a one-sided Fisher randomization test, and the tipping-points of the study are highlighted by a red contour. Historical values obtained from experts are marked on each axes.

It is evident from the displays that the study conclusion is robust to all alternative models explored, because none of the rectangular areas covered the tipping-point contour. These ETP displays reassure that there is little evidence for the differences in safety between the new kyphoplasty device and the traditional vertebroplasty procedure. Considering that analysis of the primary endpoints showed significant benefit of the new device, our TP analysis and displays helped advance the approval of the device by the FDA.



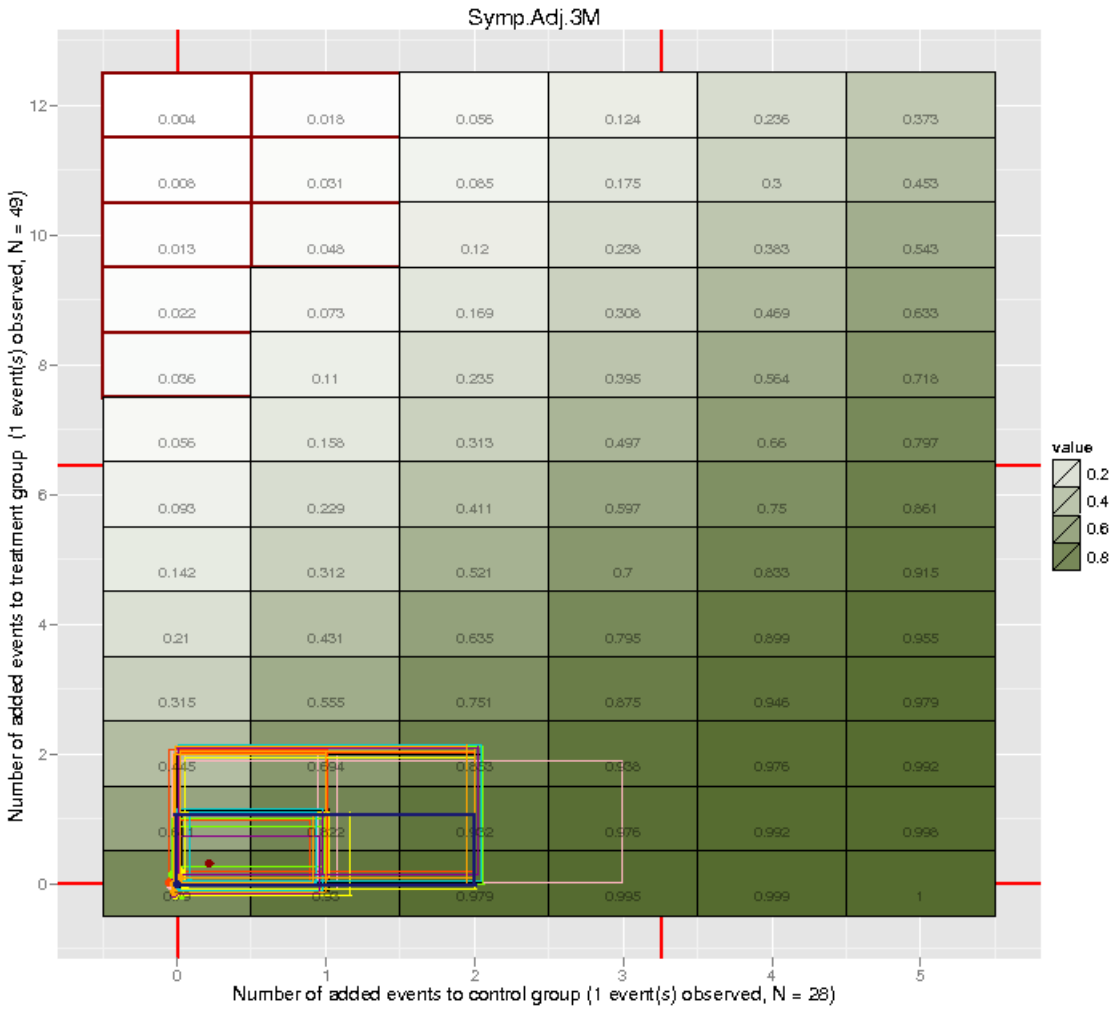
(a)

Figure 2.8: ETP displays for the twelve adverse events in the clinical trial described in Section 2.3.2, with (jittered) rectangles showing ranges of the number of successes for nonrespondents in treatment group and control group, imputed under the MAR assumption (thick blue rectangle) as well as under each of the 32 alternative models chosen for the sensitivity analysis. A vast majority of the models lead to the same conclusion of no difference in the rate of adverse events between the treatment group and the control group. Only a couple of models for the adjacent symptomatic fractures (Figure 2.8i) produced borderline imputations.



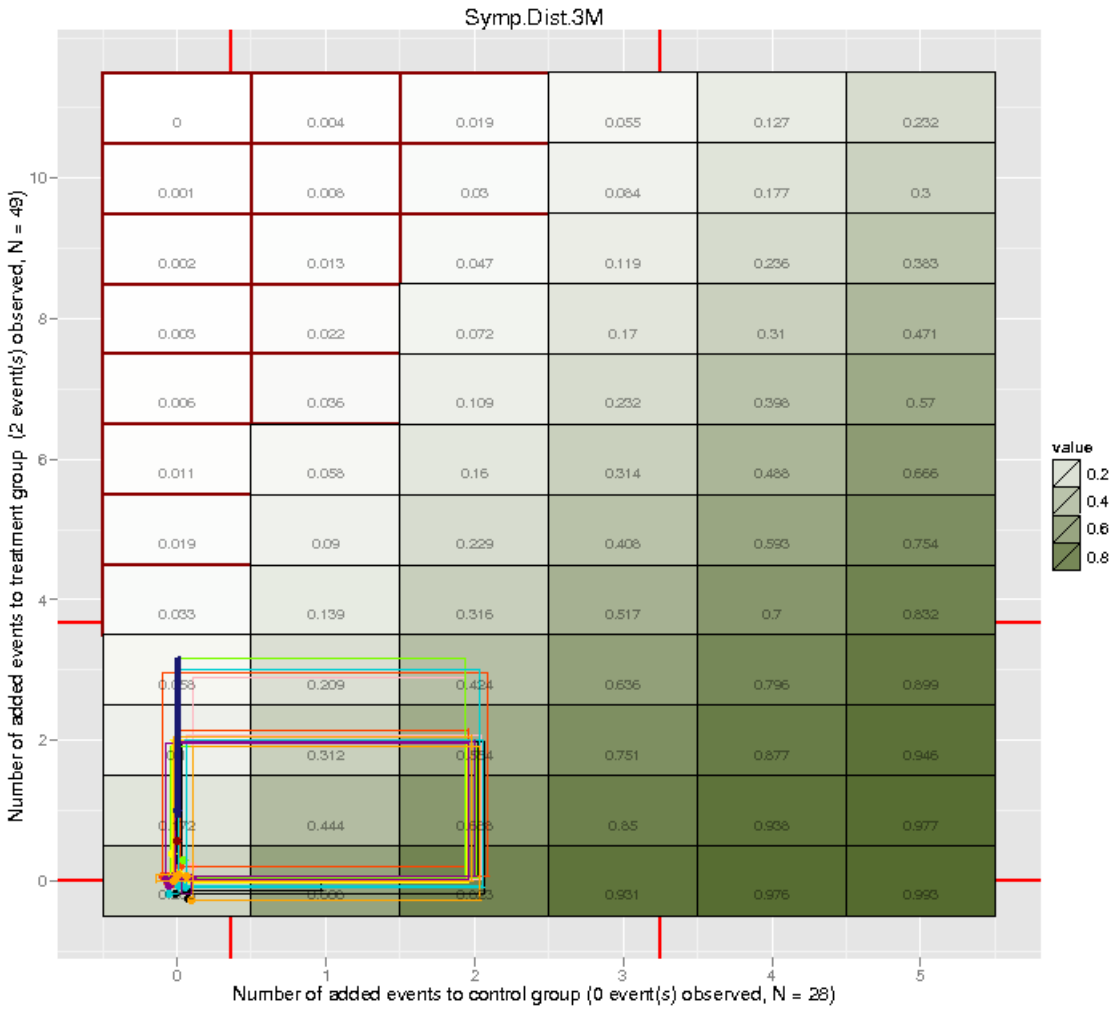
(b)

Figure 2.8: Continued.



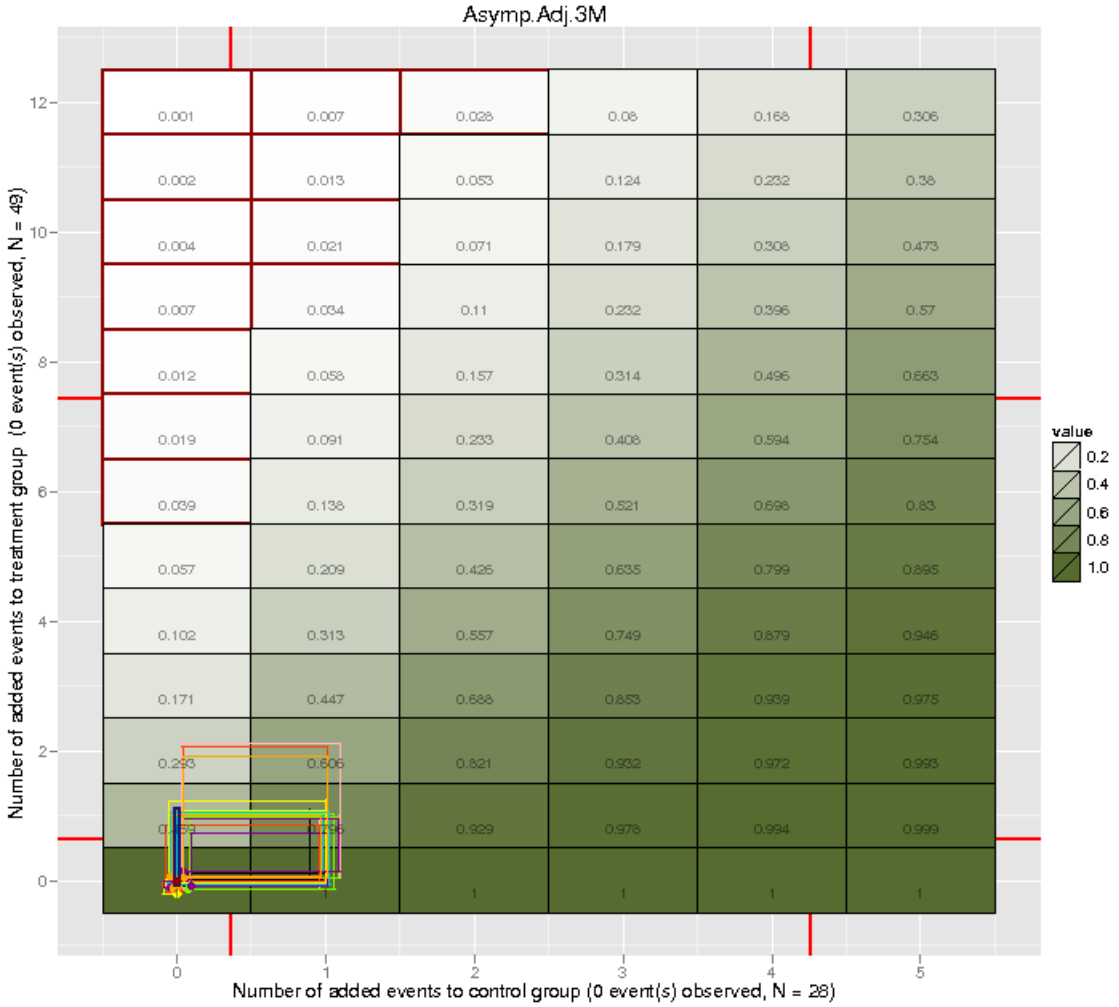
(c)

Figure 2.8: Continued.



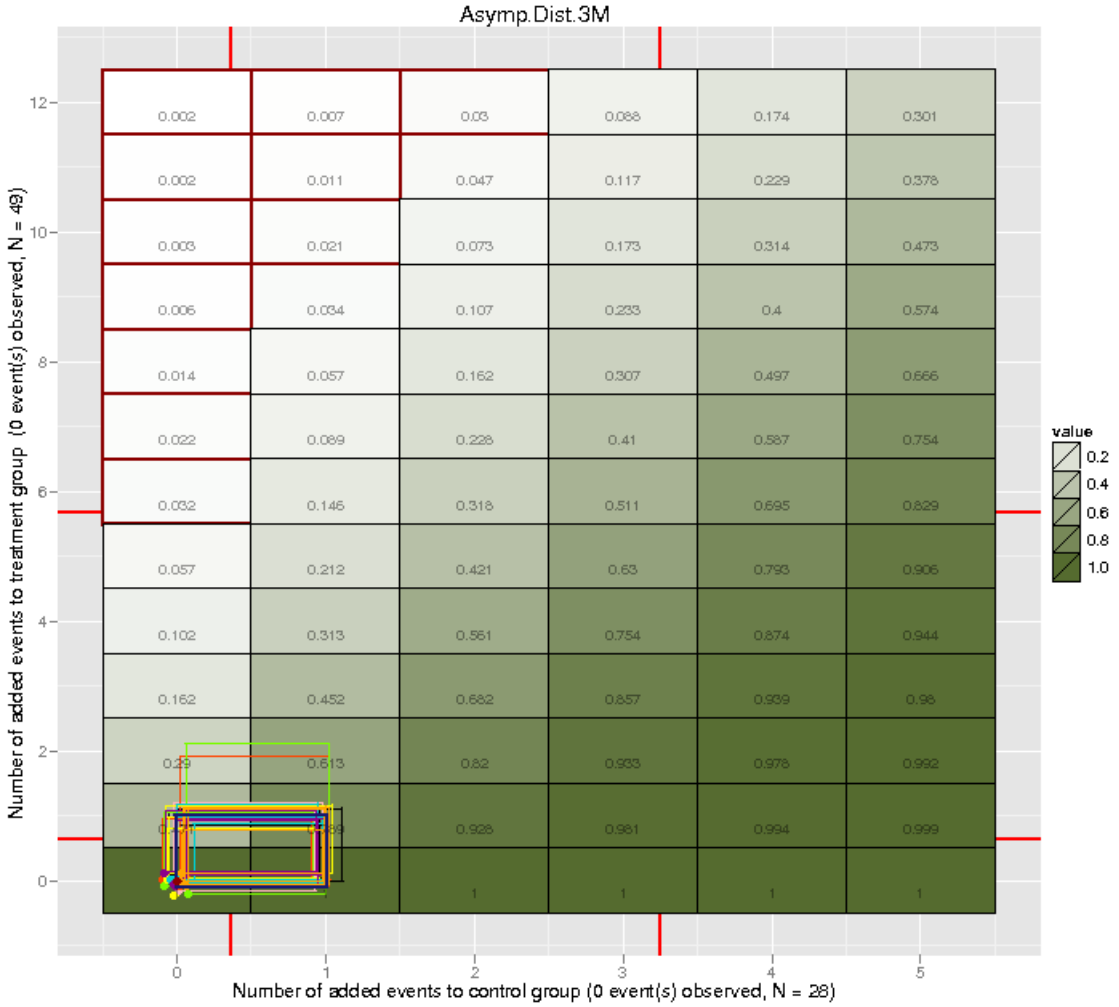
(d)

Figure 2.8: Continued.



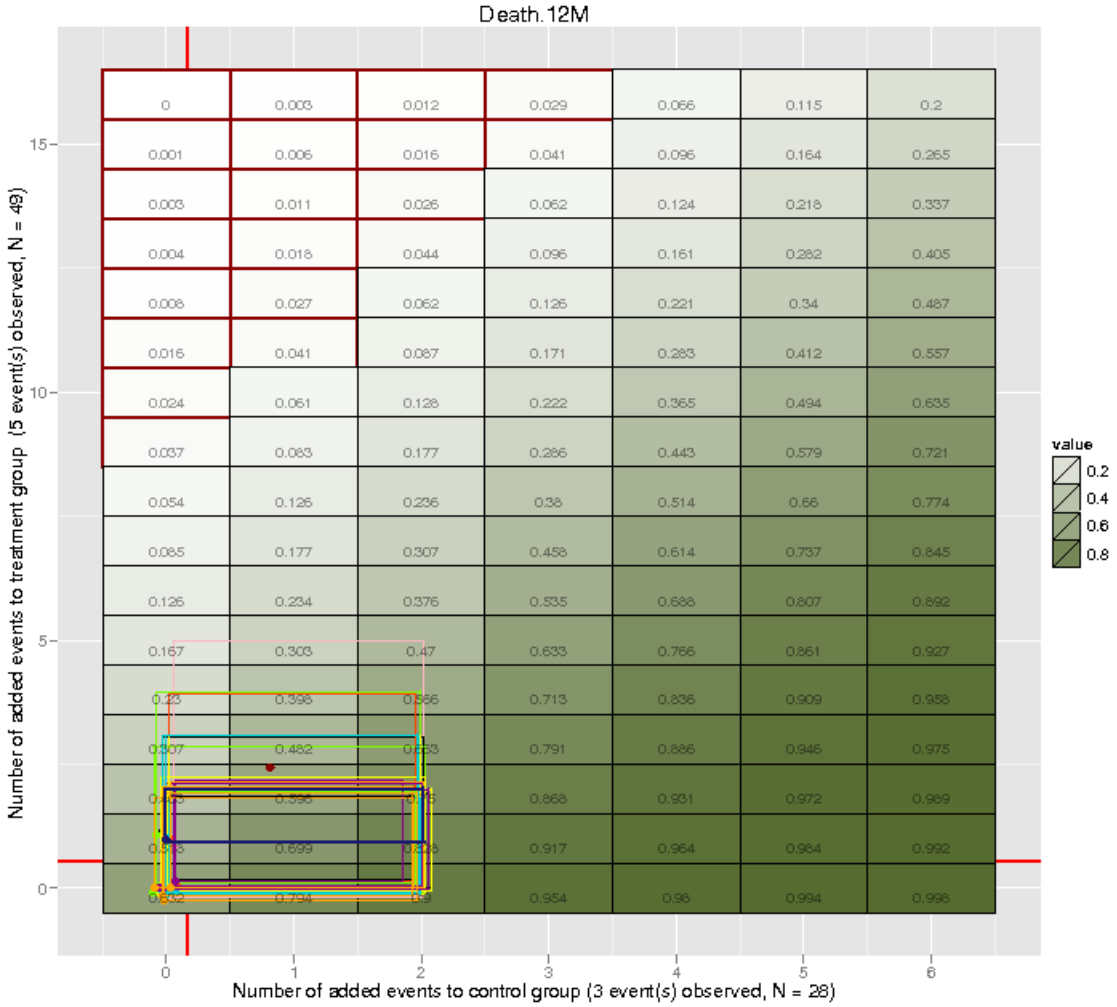
(e)

Figure 2.8: Continued.



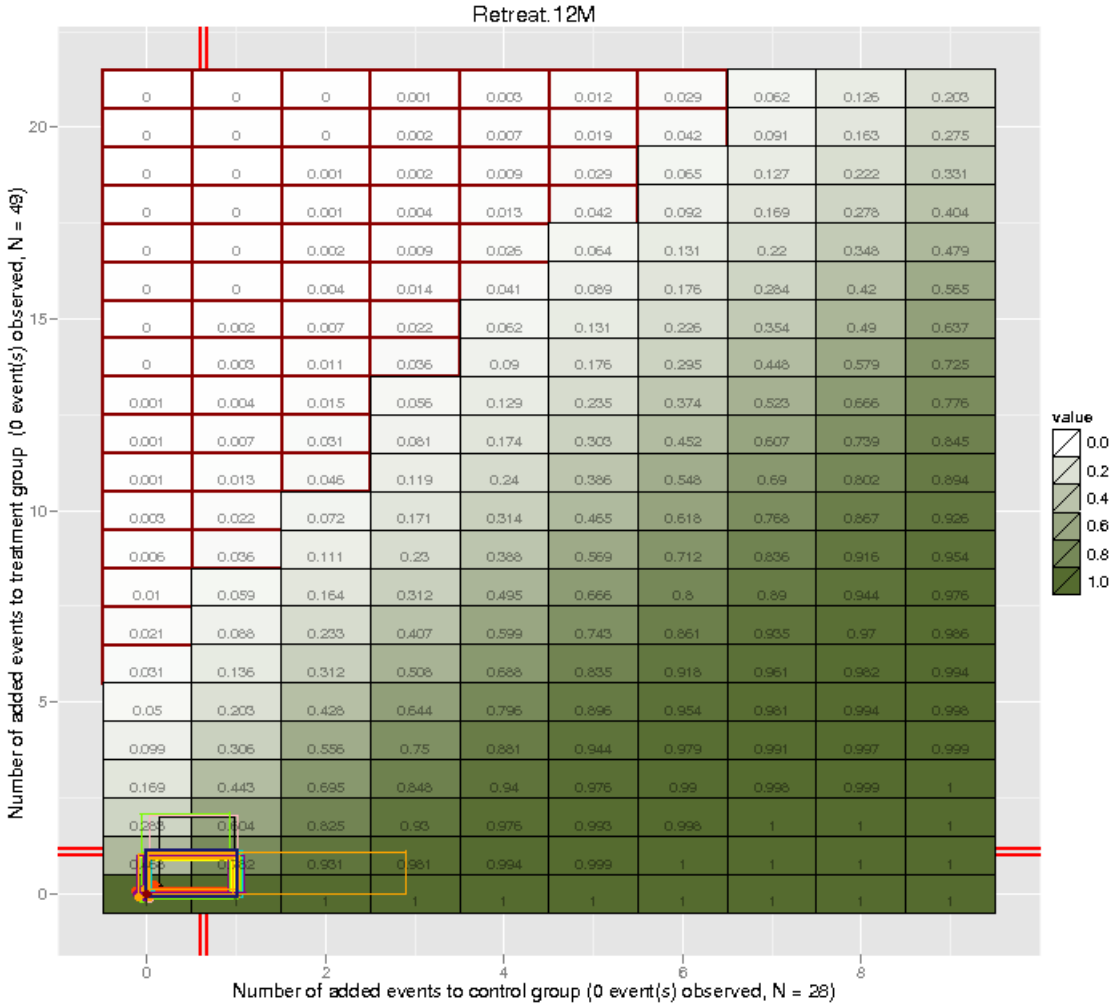
(f)

Figure 2.8: Continued.



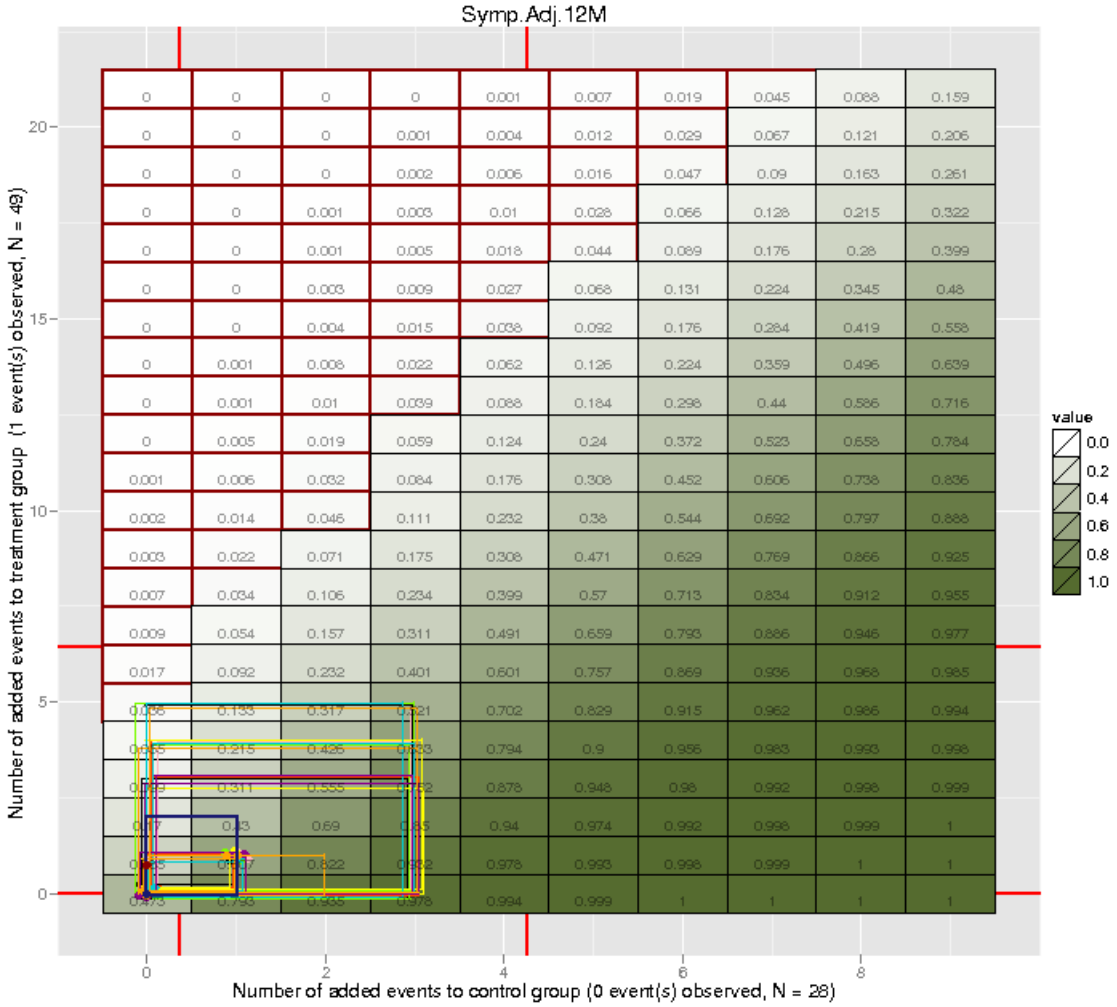
(g)

Figure 2.8: Continued.



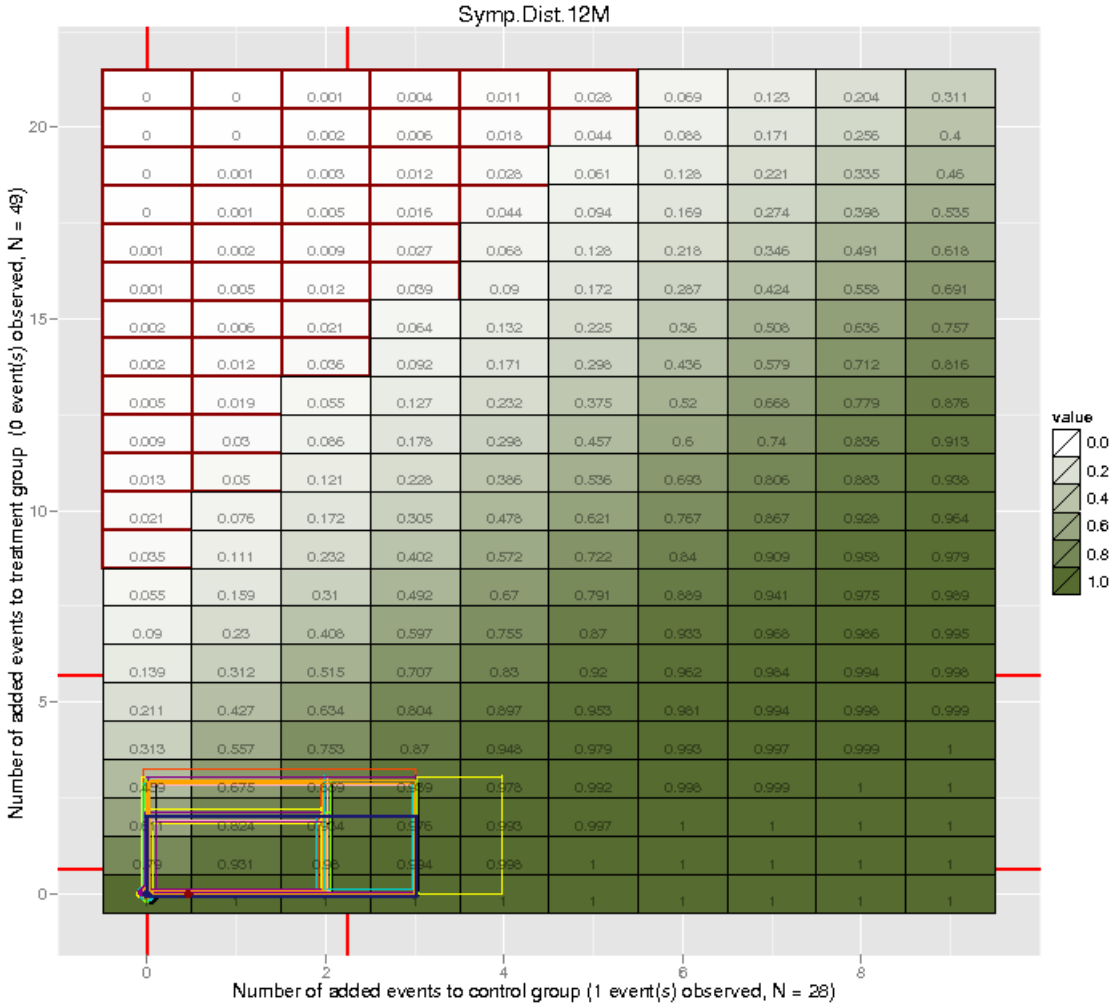
(h)

Figure 2.8: Continued.



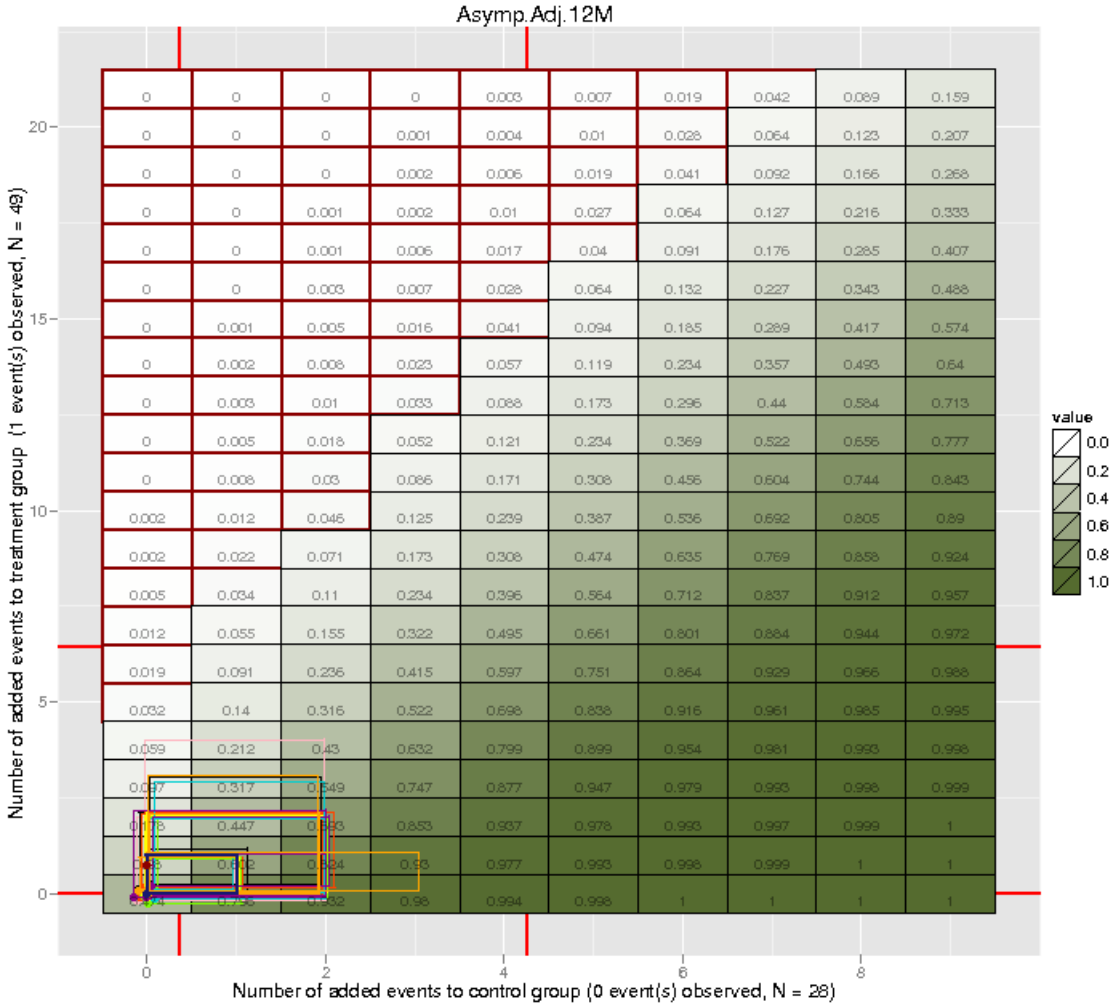
(i)

Figure 2.8: Continued.



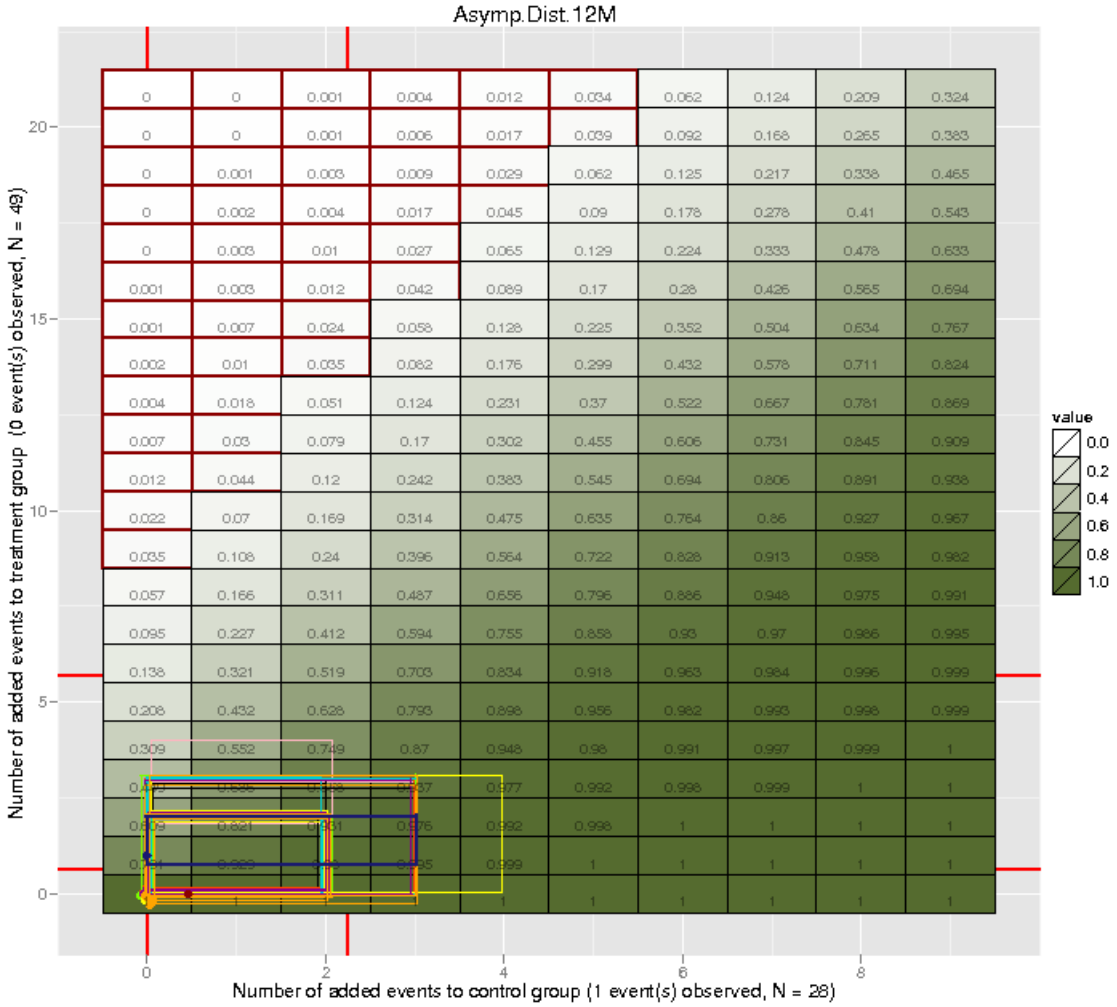
(j)

Figure 2.8: Continued.



(k)

Figure 2.8: Continued.



(l)

Figure 2.8: Continued.

2.4 Discussion

In this chapter we proposed a systematic way to perform sensitivity analyses in studies with binary outcomes, that are partially missing, using enhanced TP displays. The displays facilitate the assessment of the strength of study's conclusions under the adopted assumptions and inform us about the effect of alternative models on the conclusions. They systematize sensitivity analyses by taking advantage of modern computing to create MIs under the current and alternative models, and to display results using modern graphics.

Often, when assessing the impact of missing data on the study's conclusion, researches focus on the *worst-case scenario*, i.e., treated subjects with missing outcomes are assumed to have zero successes and, at the same time, missing outcomes for controls are set to be all favorable. In fact, in the simulated example shown in Section 2.3, this scenario would reverse the *sign* of the treatment effect, as it is evident from Figures 2.4 and 2.6. The advantage of the ETP displays is that they allow the assessment of other intermediate combinations, which are usually more realistic than the worst-case scenario. Moreover, the displays can help to convey the fact that the worst-case scenario may be unachievable, even if alternative assumptions, including MNAR, about missing data mechanism are employed.

In the real-data example in Section 2.3.2, we tackled several issues at once, including substantial missingness in the outcomes and small sample sizes in treatment and control groups. A thorough sensitivity check is a key step in this situation, exploring plausible models with alternative assumptions about the nature of missingness mechanism, including MNAR. An intuitive way to explore MNAR models is to use the

fitted outcome model under the MAR assumption as a baseline and introduce various modifications for the nonrespondents' model, informed by experts in the field. In addition, ETP displays themselves may suggest possible directions for alternative models that may tip the study's conclusions. This approach provides a new collection of useful tools for the analysis of data sets plagued with missing values. In the next chapter we generalize this idea to studies with other types of outcomes.

Chapter 3

Sensitivity Analysis using Enhanced Tipping-Point Displays for Studies with a Dichotomous Treatment and Partially Missing Outcomes.

3.1 Introduction

An assumption is considered *unassessable* if there is no statistical procedure that can be applied to available data that would support the assumption, without adding more restrictions. For example, the choice of prior distributions for a set of model parameters, the assumption of unconfounded treatment mechanism in nonrandom-

ized experiments, and the stable unit treatment value assumption (SUTVA) in causal inference are often unassessable. As we saw in Chapter 2, another area of statistics where unassessable assumptions are necessary is missing data modeling. It is recommended to perform sensitivity checks when drawing conclusions from data with missing values, especially if important policy decisions are at stake (NRC-Panel 2010; Burzykowski et al. 2010; CHMP 2010), thereby revealing how sensitive the conclusions are to the assumptions about the missing data mechanism.

In Chapter 2 we proposed a visualization, an ETP display, that allows for intuitive and systematic exploration of various MAR and MNAR assumptions, and assessment of their influence on a study's conclusions, for cases with a dichotomous treatment and a binary outcome. Here, we generalize the proposed displays to cases with continuous outcome(s), and provide a collection of distributions for which ETP displays can be easily constructed. Also, we provide a way to systematize the sensitivity analysis by introducing a set of sensitivity parameters that arise from a pattern-mixture factorization of the outcome distribution. Finally, we demonstrate how ETP displays can be utilized to guide the sensitivity analysis by suggesting *directions of high sensitivity*.

The rest of this chapter is organized as follows. Section 3.2 provides a general framework for constructing ETP displays. It includes a simulation study that demonstrates the use of ETP displays with a partially missing continuous outcome and addresses some challenges that arise for this case. In Section 3.3 we define a set of sensitivity parameters and demonstrate their use. We conclude the chapter with a description of available software packages to construct ETP displays and perform sensitivity analyses (Section 3.4) and a discussion. In addition, Appendix B.1 char-

Table 3.1: Outcome subgroups based on the treatment assignment indicator t_i and the missingness indicator d_i.

		t_i		
		0	1	
d_i	0	$\boldsymbol{Y}^C_{obs}$	$\boldsymbol{Y}^T_{obs}$	$\boldsymbol{Y}_{obs}$
	1	$\boldsymbol{Y}^C_{mis}$	$\boldsymbol{Y}^T_{mis}$	$\boldsymbol{Y}_{mis}$
		$\boldsymbol{Y}^C$	$\boldsymbol{Y}^T$	$\boldsymbol{Y}$

acterizes a set of distributions for which ETP displays can be easily built.

3.2 General Framework for ETP Displays

As before, $\boldsymbol{T} = (t_1, \ldots, t_N)$ is a vector of binary treatment indicators for N subjects. Let $\boldsymbol{Y} = (y_1, \ldots, y_N)$ be a vector of univariate outcomes (not necessarily binary) with missing values denoted by the vector of missingness indicators $\boldsymbol{D} = (d_1, \ldots, d_N)$. As in Section 2.3, given $\boldsymbol{T}$ and $\boldsymbol{D}$, the vector of outcomes $\boldsymbol{Y}$ can be partitioned into four sets, $\boldsymbol{Y}^T_{obs}, \boldsymbol{Y}^T_{mis}, \boldsymbol{Y}^C_{obs}$, or $\boldsymbol{Y}^C_{mis}$ as shown in Table 3.1. Again, suppose a goal of the study is to estimate some estimand τ, e.g., the average treatment effect, or determine a significance level for a test of τ and provide a confidence interval for it. The impact of missing values on the estimate can be illustrated by an ETP display with horizontal and vertical axes representing a function (or a *summary*) $g(\cdot)$ of values of missing outcomes for treated and control groups, $g(\boldsymbol{Y}^T_{mis})$ and $g(\boldsymbol{Y}^C_{mis})$. The analyst may choose any summary of interest as long as it is easily interpretable for the intended audience.

ETP display allows us to study some quantity of interest $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$ for each combination of $g(\boldsymbol{Y}^T_{mis})$ and $g(\boldsymbol{Y}^C_{mis})$. For example, $q(\cdot)$ could be an estimate of τ or a

p-value from a hypothesis test (i.e., t-test, noninferiority test, Fisher's randomization test etc.) used in the study. In addition, two displays can illustrate upper and lower bound of a confidence interval for the estimate of τ. Values of $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$ are illustrated on the display's background by a "heat-map", i.e., a matrix of colors, where the colors reflect the magnitude and the sign of $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$. The heat-map can be drawn by partitioning both axes into small intervals and evaluating $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$ at all possible combinations of $g(\boldsymbol{Y}^T_{mis})$ and $g(\boldsymbol{Y}^C_{mis})$ within a reasonable range.

The quantity of interest $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$ has to be a function of a pair of summaries $\{g(\boldsymbol{Y}^T_{mis}), g(\boldsymbol{Y}^C_{mis})\}$, i.e., every pair of summaries should correspond to one value of $q(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X})$, although this function can be many-to-one. Convenient choices for such summaries are *minimal sufficient statistics* (MSS) for parameter τ. When the MSS is multidimensional, we can use one component of interest, while keeping others fixed (see Section 3.2.1). Appendix B.1 provides further discussion of this approach and identifies a particular set of widely-used distributions, especially suited for ETP displays, with one-dimensional MSS readily available.

As described in Section 2.3, some supplemental information can be added to augment the sensitivity analyses. First, for any hypothesis test, the region with tipping-points can be highlighted. Second, vertical and horizontal lines can represent meaningful reference points for $g(\boldsymbol{Y}^T_{mis})$ and $g(\boldsymbol{Y}^C_{mis})$. For example, if $g(\cdot)$ represents the average outcome for nonrespondents, then lines can mark the average, minimum and maximum values observed in the data. Third, tick on axes can represent historical values of average outcomes, if available, for subjects with similar background characteristics that underwent similar treatments.

A final layer of the ETP display summarizes posterior probabilities of each of the combinations $\{g(\boldsymbol{Y}^T_{mis}), g(\boldsymbol{Y}^C_{mis})\}$ under various models for $f(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\phi})$. The posterior distributions can be calculated explicitly or approximated by means of MI. The joint distribution can be summarized on a display in several ways: contour lines, $(1-\alpha)100\%$ credible regions (Held 2004), or probability regions, approximated using Mahalanobis distance, as we illustrate in a simulated example in the next section.

3.2.1 Example with a Continuous Outcome

Consider a study with N subjects, randomly divided between treatment and control groups, treatment assignment vector $\boldsymbol{T}$ and outcome values $\boldsymbol{Y}$. Suppose the outcome is blood pressure, measured for each subject post-treatment. In addition, two baseline fully-observed predictors are available: sex ($\boldsymbol{X}_F$, with "F" for female) and years of school ($\boldsymbol{X}_S$, 0 through 21). Some subjects are missing the outcome, as indicated by $\boldsymbol{D}$, and our task is to estimate the marginal population treatment effect on blood pressure and check its sensitivity to various assumptions about the missing data.

Vector $\boldsymbol{Y}$ has four parts, described in Table 3.1. As in Section 2.3, we let τ be the marginal average treatment effect, constant across all subjects, with its unbiased estimator given by (2.1). Natural and easily interpretable summaries of missing outcomes that can be used for axes of the ETP display in a continuous-outcome case are average responses among nonrespondents in the treatment group and the control

group,

$$g(\boldsymbol{Y}^T_{mis}) = \sum_{i:y_i \in \boldsymbol{Y}^T_{mis}} y_i/N^T_{mis} = \bar{y}^T_{mis},\ g(\boldsymbol{Y}^C_{mis}) = \sum_{i:y_i \in \boldsymbol{Y}^C_{mis}} y_i/N^C_{mis} = \bar{y}^C_{mis}.$$

Then, for a set of observed outcomes $\boldsymbol{Y}^T_{obs}$ and $\boldsymbol{Y}^C_{obs}$, $\hat{\tau}$ can be represented as follows,

$$\hat{\tau} = \frac{\bar{y}^T_{obs}N^T_{obs} + g(\boldsymbol{Y}^T_{mis})N^T_{mis}}{N^T} - \frac{\bar{y}^C_{obs}N^C_{obs} + g(\boldsymbol{Y}^C_{mis})N^C_{mis}}{N^C}. \tag{3.1}$$

The continuous nature of the response makes it more challenging to generalize ETP displays for this problem due to a wide variety of continuous distributions available for modeling $\boldsymbol{Y}$. However, there are fundamental reasons for considering the sample mean as a convenient statistic for this problem. When $\boldsymbol{Y}$ has a distribution that is a member of a *natural exponential family* of order one (NEF_1, Morris 1982, 1983), it can be shown that sample mean is the MSS. Moreover, for certain multiparameter exponential families, sample mean can also be a component of a multidimensional MSS. We describe a family of distributions that is particularly suitable for ETP displays in Appendix B.1.

In order to illustrate the application of ETP displays to the case with continuous outcomes, we use the following model to generate the data, independently for N units,

$$\begin{aligned}
&y_i = 125 + 4t_i + 0.3x_{i,S} + x_{i,F} + \epsilon_i, \text{ where}\\
&\epsilon_i \mid t_i = 0 \sim \text{N}(0, \sigma^2_C) \text{ and } \epsilon_i \mid t_i = 1 \sim \text{N}(0, \sigma^2_T),\\
&d_i \mid p_i \sim \text{Binom}(p_i), \text{ where}\\
&\text{logit}(p_i) = 7 - 0.6x_{i,S} - t_i + 0.0005y_i,\ i = 1, \ldots, N.
\end{aligned}$$

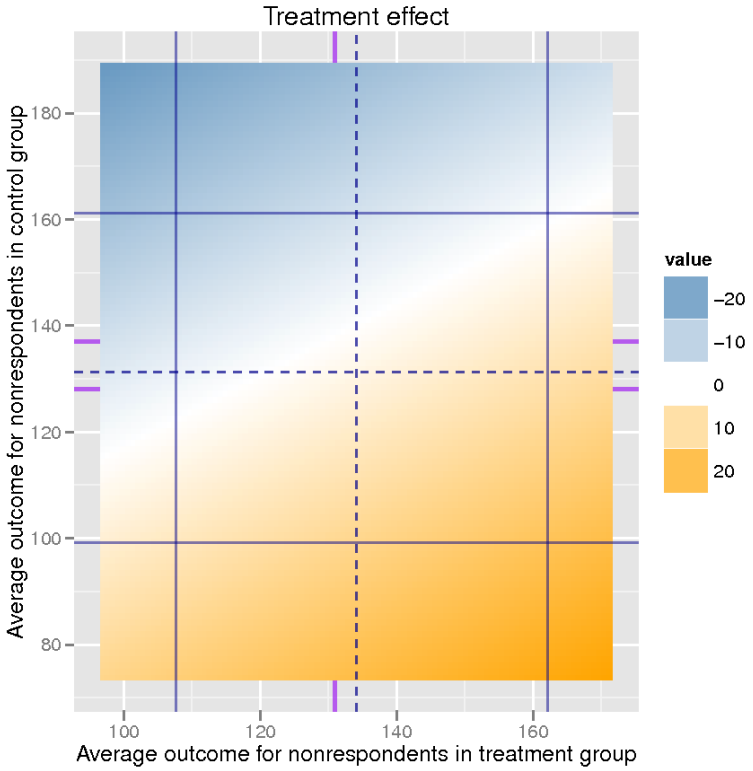


Figure 3.1: ETP display for the continuous outcome, showing estimated average treatment effects (3.1) using a heat-map. Horizontal and vertical axes represent average outcomes among nonrespondents in the treated and control groups, respectively. Two pairs of vertical and horizontal blue lines correspond to minimum and maximum values of outcomes observed in each group, and dashed blue lines represent average outcomes, 134.1 and 131.3, for respondents in treated and control groups, respectively. In addition, several horizontal and vertical ticks give historical values of the average outcome for treated and control groups that may be available to the analyst.

Thus, each outcome y_i is Normally distributed given $x_{i,S}$, $x_{i,F}$, and t_i, with different standard deviations for treated ($\sigma_T = 15$) and control ($\sigma_C = 10$) subjects. Predictors $x_{i,F}$ were generated from a Bern(0.5), and predictors $x_{i,S}$ were generated using a multinomial distribution to draw a number of years of school from $0, 1, \ldots 21$ according to a plausible vector of probabilities. Under the assumption that the pa-

rameters are unknown in each group and that the covariates are fixed, the distribution of y_i, given t_i, is an exponential family (EF), and the data has a two-dimensional MSS in each treatment group,

$$\left(\bar{y}^T, \hat{\sigma}_T^2\right) \text{ and } \left(\bar{y}^C, \hat{\sigma}_C^2\right),$$

where $\hat{\sigma}_T^2$ and $\hat{\sigma}_C^2$ are sample variances of outcomes in the treatment group and the control group, respectively. In the generated data, the outcomes were missing for 28% of the control subjects and 21% of the treated subjects. Figure 3.1 shows a heat-map of estimated treatment effects $\hat{\tau}$, calculated using (3.1), for the simulated data set.

In order to perform a hypothesis test of the null-hypothesis $H_0 : \tau = 0$, we can use a Welch's t-test. However, the test statistic for the Welch's t-test,

$$\frac{\bar{y}^T - \bar{y}^C}{\sqrt{\frac{\hat{\sigma}_T^2}{N^T} + \frac{\hat{\sigma}_C^2}{N^C}}}, \tag{3.2}$$

depends on sample means and sample variances of all outcomes, including the missing ones. Therefore, in order to represent the results of the test on the ETP display, we need to make additional assumptions.

Theorem 3.2.1. *Suppose $y_1, \ldots, y_K \sim Norm(\mu, \sigma^2)$ with K_{obs} values observed and K_{mis} values missing completely at random, $K = K_{mis} + K_{obs}$. Suppose that we also know the sample average of missing values, $\bar{y}_{mis}$. Then the uniformly minimum-variance unbiased estimator (UMVUEs) of μ and σ^2 will be*

$$\hat{\mu} = \frac{K_{obs}\bar{y}_{obs} + K_{mis}\bar{y}_{mis}}{K} \text{ and } s^2 = \frac{(K_{obs} - 1)\hat{\sigma}_{obs}^2 + \frac{K_{obs}K_{mis}}{K}(\bar{y}_{obs} - \bar{y}_{mis})^2}{K_{obs}}, \tag{3.3}$$

where $\bar{y}_{obs}$, $\bar{y}_{mis}$ *are sample means of observed and missing values, and* $\hat{\sigma}^2_{obs}$ *is the sample variance of the observed values. Also,* $\hat{\mu}|\mu, \sigma^2 \sim Norm(\mu, \sigma^2/K)$ *and* $K_{obs}s^2|\sigma^2 \sim \sigma^2\chi^2_{K_{obs}}$.

Proof. It is easy to show that both estimators are unbiased, and, because they are based on complete sufficient statistics, according to the Lehmann-Scheffe theorem, they are UMVUEs. According to the Basu's theorem, with respect to the parameter μ, the complete sufficient statistic $\bar{y}_{obs}$ and the ancillary statistic $\hat{\sigma}^2_{obs}$ are independent, and both are also independent from $\bar{y}_{mis}$, because the missingness is MCAR. Sampling distributions of $\hat{\mu}$ and s^2 are evident after we recognize that

$$\bar{y}_{obs} - \bar{y}_{mis} \sim Norm\left(0, \sigma^2\left(\frac{1}{K_{obs}} + \frac{1}{K_{mis}}\right)\right)$$

□

Theorem 3.2.1 provides a way to construct a pivot, analogous to (3.2).

Theorem 3.2.2. *The approximate sampling distribution of*

$$\frac{\hat{\mu}^T - \hat{\mu}^C}{\sqrt{\frac{s_1^2}{N^T} + \frac{s_0^2}{N^C}}}, \tag{3.4}$$

where $\hat{\mu}^T$ *and* s_1^2 *are calculated according to (3.3) for the treated subjects, and* $\hat{\mu}^C$ *and* s_0^2 *are calculated similarly for the controls, is a* t*-distribution with the following degrees of freedom*

$$\hat{\hat{f}} = \frac{(s_1^2/N^T + s_0^2/N^C)^2}{(s_1^2/N^T)^2/N^T_{obs} + (s_0^2/N^C)^2/N^C_{obs}}. \tag{3.5}$$

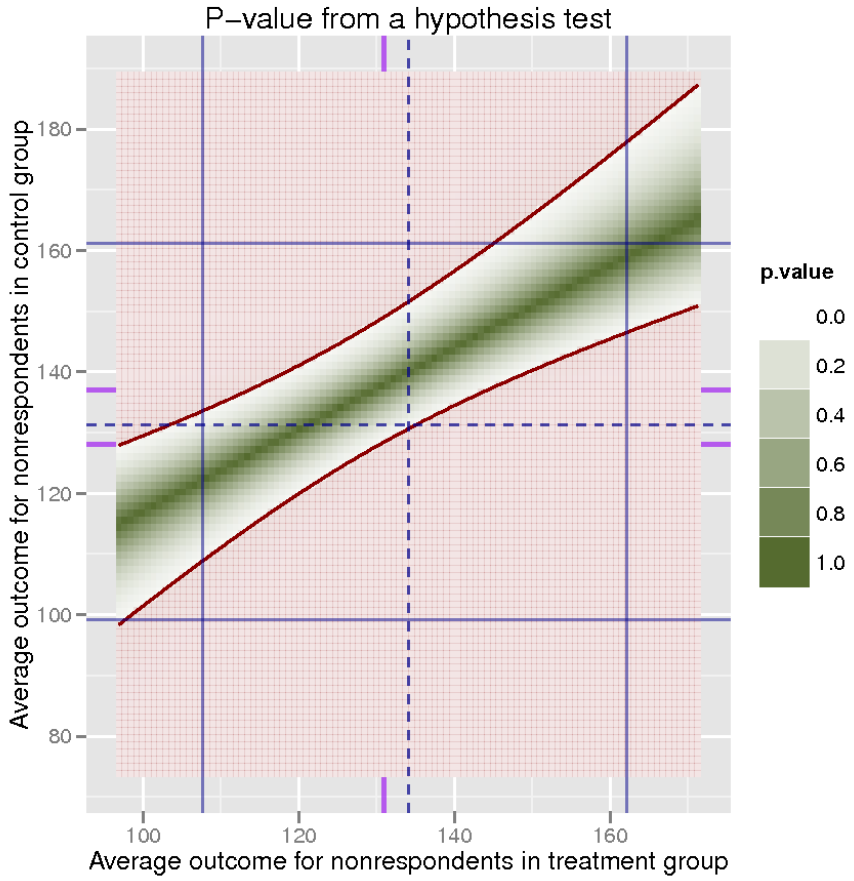


Figure 3.2: ETP display for a continuous outcome, displaying p-values from a two-sided Welch's t-test. For each combination of average outcomes for nonrespondents in the treatment group and the control group, the test-statistic is calculated according to (3.4). The red contour highlights tipping-points of the study, which correspond to the 0.05 significance level, and the heat-map represents the magnitude of the p-values.

The proof, similar to the one in Welch (1938), is presented in Appendix B.2. Figure 3.2 shows the heat-map of p-values, obtained form the derived Welch t-test, with a tipping-point contour that corresponds to a significance level of 0.05.

Finally, we proceed with analyzing the data by multiply imputing missing values and estimating the population treatment effect under MCAR and MAR assumptions. Figure 3.3 shows two convex hulls that contain 95% of average outcome values for

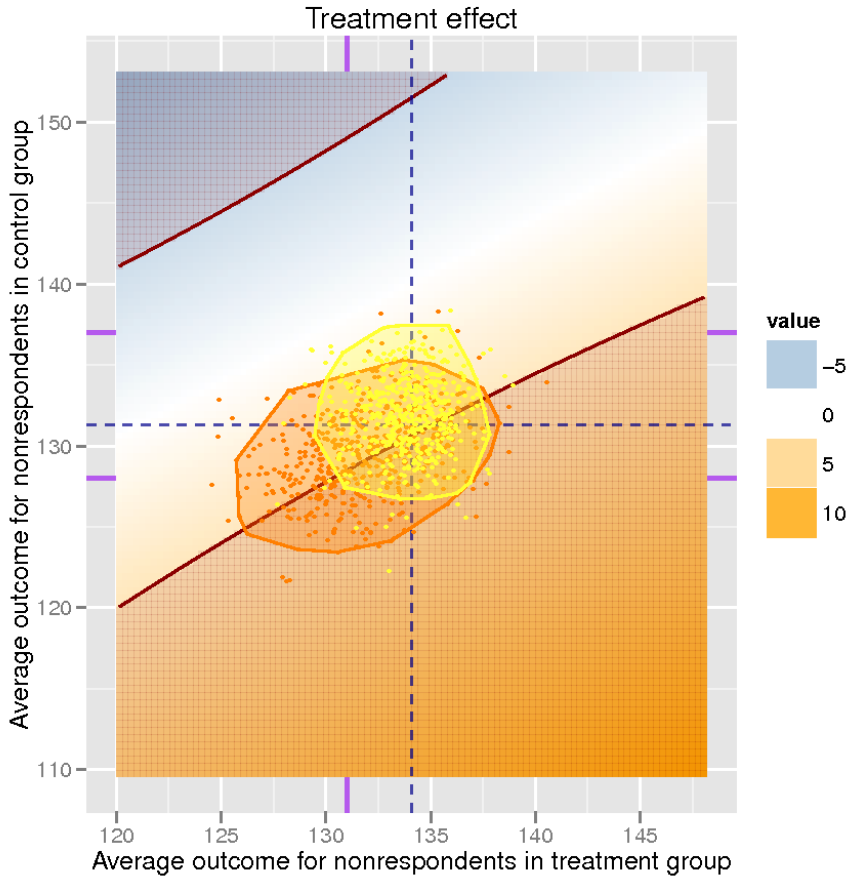


Figure 3.3: ETP display with two convex hulls, each containing 95% of the 500 MIs, generated under the MCAR (yellow) and the MAR (orange) assumptions. The excluded 5% of points have the largest Mahalanobis distance from the sample mean. Background colors correspond to the estimates of the treatment effect, and red-shaded region identifies combinations that would result in a "significant" treatment effect, according to the derived two-sided Welch t-test, at the traditional 0.05 level. Note that MAR and MCAR assumptions result in different sets of imputations, however, neither of the two models give a clear answer to the question concerning a treatment effect.

nonrespondents in treatment and control groups, produced by the MI procedure, excluding 5% of imputations with the largest Mahalanobis distance from the sample mean. The hulls approximate 95% posterior regions of joint distributions of average outcomes among nonrespondents in the treatment and control groups under each

Table 3.2: Treatment effect on $\boldsymbol{Y}$, estimated under MAR and MCAR assumptions by combining results from 500 MIs using Rubin's rule.

MI assumption	Estimated treatment effect	95% posterior interval
MCAR	2.6	(-0.85, 6.00)
MAR	2.8	(-0.60, 6.20)

model.

Note that neither of the two models produces conclusive results. Although, both of them result in estimates of the treatment effect that are not significant (see Table 3.2), Figure 3.3 reveals how sensitive the results are to the assumptions about the missingness mechanism. In the next section we propose a systematic way to formulate alternative MAR and MNAR assumptions, and to utilize ETP displays to study the sensitivity of the treatment effect estimate.

3.3 Exploring MNAR models with ETP displays

The majority of methods developed to handle missing data require MAR assumption, and there is a shortage of standardized ways to explore alternative assumptions systematically. Intuitively, there are two situations that violate MAR: (1) there exists an unobserved (or "lurking") variable $\boldsymbol{U}$ that is an important predictor for both, the partially-missing outcome and the missingness, or (2) the missingness mechanism depends on the unobserved outcome itself. For example, suppose study subjects are more likely to miss a follow-up appointment if they develop a complication after the treatment, recorded by the variable $\boldsymbol{U}$. At the same time, the outcome of interest $\boldsymbol{Y}$, e.g., subject's pain level, also depends on whether any complication has occurred. Therefore, failing to record $\boldsymbol{U}$ will result in the missingness that is not MAR. More-

over, this situation also leads to the omission of the important predictor of $\boldsymbol{Y}$ from the model $f(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$.

On the other hand, if we assume that all predictors that affect $\boldsymbol{Y}$ and $\boldsymbol{D}$ are collected, then the only situation that leads to MNAR is when the missingness in $\boldsymbol{Y}$ depends on unobserved values. In our example, this would mean that the pain level itself made it more difficult for subjects to attend the follow-up appointment. An important practical difference between the two situations described above is that the first one can be avoided by careful study planning to anticipate the reasons for, and minimize, dropouts (as recommended in NRC-Panel 2010, Ch.2-3), and the second one can only be handled at the analysis stage, by modeling the missingness mechanism explicitly and, thus, introducing more unassessable assumptions.

Recently, the NRC-Panel (2010) described a basic procedure that can be used to systematize sensitivity analyses for experiments with missing data. The procedure was based on the idea, originally proposed in Rubin (1977), of constructing the outcome distribution for nonrespondents using the outcome distribution estimated for respondents under the MAR assumption, but distorting it in a systematic manner. Here we show how ETP displays can assist in a systematic exploration of alternative distributions for the outcome. In particular, they can help identify deviations from the MAR model along *directions of high sensitivity*, i.e., the types of models for the nonrespondents' outcomes in treatment and control groups that are likely to change the study's conclusions.

Suppose that the study units are independent and exchangeable. Parametric inference for incomplete data with MNAR missingness requires specification of the

joint distribution of the outcome and the missingness mechanism, $f(y, d \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$, identical for all units. Two approaches to modeling MNAR mechanism data can be used: selection models and pattern-mixture models. *Selection models* (Rubin 1974; Heckman 1976) are based on the following factorization of the joint distribution,

$$f(y, d \mid \boldsymbol{x}; \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}}) = f(d \mid y, \boldsymbol{x}; \tilde{\boldsymbol{\phi}}) f(y \mid \boldsymbol{x}; \tilde{\boldsymbol{\theta}}),$$

with $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\theta}}$ a priori independent. This approach uses the idea of weighting the marginal distribution of the outcome $f(y \mid \boldsymbol{x}; \tilde{\boldsymbol{\theta}})$ by a selection probability, $f(d \mid y, \boldsymbol{x}; \tilde{\boldsymbol{\phi}})$, that accounts for a nonrandom nonresponse. Both parts of the factorization have to be postulated, because there is no way to estimate either of them empirically from data alone under the MNAR assumption.

Another approach called *pattern-mixture models* (Little 1994), first introduced in Rubin (1977) and further pursued in Glynn et al. (1986), arises from a different factorization,

$$\begin{aligned} f(y, d \mid \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) &= f(y \mid d, \boldsymbol{x}; \boldsymbol{\theta}) f(d \mid \boldsymbol{x}; \boldsymbol{\phi}) \\ &= \begin{cases} f(y \mid d = 0, \boldsymbol{x}; \boldsymbol{\theta}) P(d = 0 \mid \boldsymbol{x}; \boldsymbol{\phi}) & \text{if } d = 0, \\ f(y \mid d = 1, \boldsymbol{x}; \boldsymbol{\theta}) P(d = 1 \mid \boldsymbol{x}; \boldsymbol{\phi}) & \text{if } d = 1, \end{cases} \end{aligned} \tag{3.6}$$

with $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ a priori independent. Under MAR, $f(y \mid d = 0, \boldsymbol{x}; \boldsymbol{\theta}) = f(y \mid d = 1, \boldsymbol{x}; \boldsymbol{\theta})$ but, in general, the outcome models for respondents and nonrespondents may differ, and the joint model for y and d is a mixture of the two models, $f(y \mid d = 0, \boldsymbol{x}; \boldsymbol{\theta})$ and $f(y \mid d = 1, \boldsymbol{x}; \boldsymbol{\theta})$.

An immediate benefit of the factorization in (3.6) is that two out of three components, $f(d \mid \boldsymbol{x};\boldsymbol{\phi})$ and $f(y \mid d=0,\boldsymbol{x};\boldsymbol{\theta})$, can be estimated from the observed data, and the only part that requires unverifiable modeling assumptions is the conditional distribution of the outcomes for nonrespondents, i.e., $f(y \mid d=1,\boldsymbol{x};\boldsymbol{\theta})$. Another advantage of pattern-mixture models is that they allow for a natural formulation (and interpretation) of alternative models for the purpose of performing sensitivity analyses. Because the outcome distributions for respondents and nonrespondents are specified separately, alternative models for nonrespondents can be formed by introducing various deviations to the model estimated for respondents. Moreover, the nature of these deviations can be discussed with experts in the field. The two approaches, selection and pattern-mixture modeling, are fundamentally related, as it is demonstrated in Buuren (2012, Sec. 3.9.4) using Bayes Rule; also see extensive exchanges on this topic by discussants in Wainer (1986).

Table 3.3 systematizes some types of modifications that can be used to model anticipated differences between the distributions of outcomes for respondents and nonrespondents when the outcome is continuous. Each row in Table 3.3 introduces a new *sensitivity parameter* that can affect the study's conclusion. Types 1 through 4 focus specifically on changes in the average outcome, types 5 and 6 modify the variance of the outcome, and types 7 and 8 modify the entire distribution. These types may be used one-by-one, as well as in any combination. For example, a family of modifications that link the marginal response for respondents to the one for nonrespondents can be represented by a set of sensitivity parameters $\{\delta, \beta_{x_1}, \ldots, \beta_{x_K}, \upsilon\}$. It implies that the expected response for nonrespondents, $E(y \mid \boldsymbol{x}, d=1,\boldsymbol{\theta})$, is different from the

Table 3.3: Types of sensitivity parameters that can be introduced in order to link the distribution of outcomes for respondents $f(y \mid \boldsymbol{x}, d = 0; \boldsymbol{\theta})$ and nonrespondents $f(y \mid \boldsymbol{x}, d = 1; \boldsymbol{\theta})$ for a continuous outcome y. Here, χ is a proper subspace of the covariate space, and $\boldsymbol{x} \in \chi$.

Type	Modification	Description
1	$E(y \mid d = 1; \boldsymbol{\theta}) = E(y \mid d = 0; \boldsymbol{\theta}) + \delta$	Marginal mean response shift
2	$E(y \mid \boldsymbol{x}, d = 1; \boldsymbol{\theta}) = E(y \mid \boldsymbol{x}, d = 0; \boldsymbol{\theta}) + \delta_{\chi}$	Conditional mean response shift
3	$E(y \mid d = 1; \boldsymbol{\theta}) = E(y \mid d = 0; \boldsymbol{\theta}) + \beta_{x_j} x_j$	Marginal effect change for x_j
4	$E(y \mid \boldsymbol{x}, d = 1; \boldsymbol{\theta}) = E(y \mid \boldsymbol{x}, d = 0; \boldsymbol{\theta}) + \beta_{\chi, x_j} x_j$	Conditional effect change for x_j
5	$Var(y \mid d = 1, \boldsymbol{\theta}) = \upsilon Var(y \mid d = 0, \boldsymbol{\theta})$	Marginal variance scaling
6	$Var(y \mid \boldsymbol{x}, d = 1, \boldsymbol{\theta}) = \upsilon_{\chi} Var(y \mid \boldsymbol{x}, d = 0, \boldsymbol{\theta})$	Conditional variance scaling
7	$f(y \mid d = 1; \boldsymbol{\theta}) = f(y/\omega \mid d = 0; \boldsymbol{\theta})/\omega$	Response scale adjustment
8	$f(y \mid d = 1; \boldsymbol{\theta}) = f\left(g^{-1}(y) \mid d = 0; \boldsymbol{\theta}\right) \left\lvert \frac{d}{dy} g^{-1}(y) \right\rvert$	Response shape adjustment, using the transformation $g(\cdot)$

one that would have been observed if they were respondents with the same values of covariates, $E(y \mid \boldsymbol{x}, d = 0, \boldsymbol{\theta})$, by $\delta + \beta_{x_1} x_1 + \cdots + \beta_{x_K} x_K$, and its variance is different by υ. Analogously, *conditional* shifts, scaling or effect changes (types 2, 4 and 6) are modifications that affect only specific subgroups of nonrespondents. They can be introduced not only based on background characteristics, but also on the treatment arm membership, e.g., mean outcome for respondents and nonrespondents can differ by $\delta_{t=0}$ for controls only, etc.

Parameters given in Table 3.3 suggest systematic ways to perform sensitivity analyses for studies with missing data and, together with the ETP display, may reveal models that exhibit especially high sensitivity. For example, the ETP display for

the continuous case presented in Section 3.2.1 (Figure 3.3) suggests that models that reduce mean outcome for nonrespondents in the control group or increase mean outcome for nonrespondents in the treatment group, or both, can alter the conclusion of the study that the treatment effect is insignificant at the 0.05 significance level, reached under both MCAR and MAR models. The following modifications to the MAR model, or their combinations, have these properties:

- $E(y \mid d = 1, t, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t, x_F, x_S, \boldsymbol{\theta}) + \delta$;
- $E(y \mid d = 1, t = 1, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 1, x_F, x_S, \boldsymbol{\theta}) + \delta_{t=1}$, with $\delta_{t=1} > 0$;
- $E(y \mid d = 1, t = 0, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 0, x_F, x_S, \boldsymbol{\theta}) - \delta_{t=0}$, with $\delta_{t=0} > 0$;
- $E(y \mid d = 1, t = 1, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 1, x_F, x_S, \boldsymbol{\theta}) + \beta_{t=1,x_S} x_S$, with $\beta_{t=1,x_S} > 0$;
- $E(y \mid d = 1, t = 0, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 0, x_F, x_S, \boldsymbol{\theta}) - \beta_{t=0,x_S} x_S$, with $\beta_{t=0,x_S} > 0$;
- $E(y \mid d = 1, t = 1, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 1, x_F, x_S, \boldsymbol{\theta}) + \beta_{t=1,x_F} x_F$, with $\beta_{t=1,x_F} > 0$;
- $E(y \mid d = 1, t = 0, x_F, x_S, \boldsymbol{\theta}) = E(y \mid d = 0, t = 0, x_F, x_S, \boldsymbol{\theta}) - \beta_{t=0,x_F} x_F$, with $\beta_{t=0,x_F} > 0$.

After identifying combinations of parameters δ, $\delta_{t=0}$, $\delta_{t=1}$, $\beta_{t=0,x_S}$, $\beta_{t=1,x_S}$, $\beta_{t=0,x_F}$, and $\beta_{t=1,X_F}$ that change the study's conclusion, the analyst may defer to experts for deciding whether any of these combinations represent plausible alternative models. Figure 3.4 shows ETP displays with four contours that contain 95% of imputed means

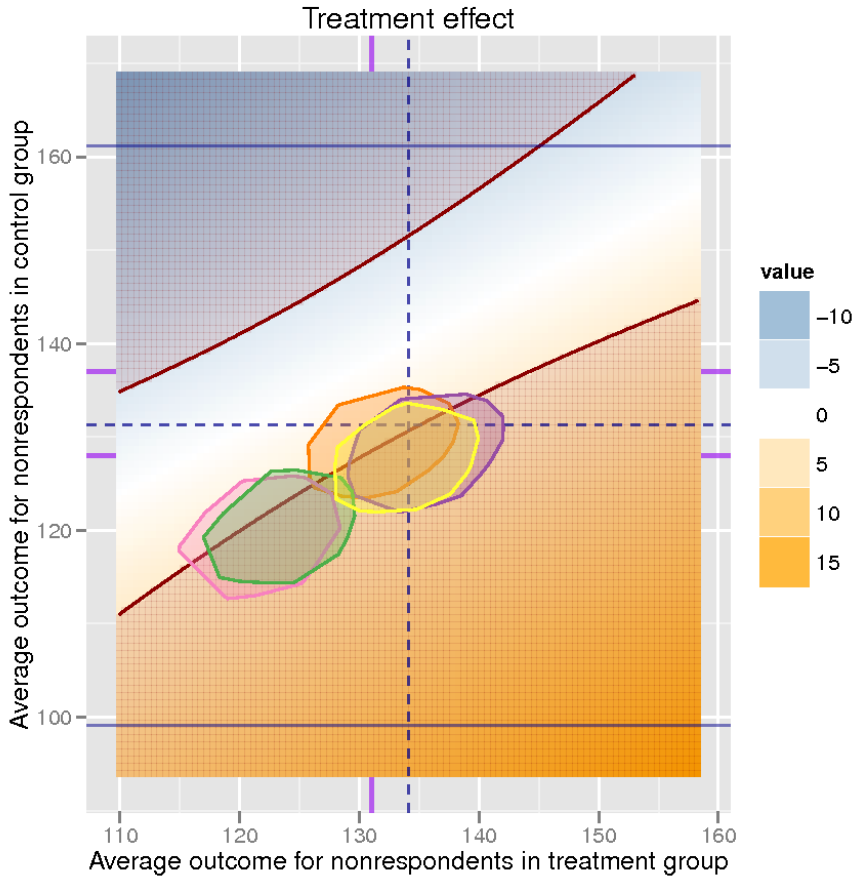


Figure 3.4: ETP display for the continuous outcome introduced in Section 3.2.1, with convex hulls that include 95% of MIs produced under four alternative MNAR models described in Table 3.4. The corresponding colors are 1-pink, 2-purple, 3-green, and 4-yellow. The orange hull includes MIs produced under the MAR model, used in Section 3.2.1. The individual imputations are not displayed. All four contours for alternative models are located in directions of high sensitivity from the contour obtained under the MAR assumption.

for nonrespondents in treatment and control groups, produced using four alternative MNAR models, along with the MIs obtained under the MAR model previously. Table 3.4 describes the models used and reports the estimated treatment effects under each of them. All four models resulted in a "significant" treatment effect, estimated closer to the true value. Therefore, if the changes introduced in any of these models

Table 3.4: Average treatment effect on the continuous outcome $\boldsymbol{Y}$, estimated under four alternative MNAR models by combining 500 MIs produced under each model, with true effect $\tau = 4$.

Model	$\{\delta, \delta_{t=0}, \delta_{t=1}, \beta_{t=0,x_S}, \beta_{t=1,x_S}, \beta_{t=0,x_F}, \beta_{t=1,x_F}\}$	Estimated average treatment effect	95% posterior interval
1	$\{-10, 0, 0, 0, 0, 0, 0\}$	3.91	$(0.26, 7.56)$
2	$\{0, 1, 4, 0, 0, 0, 0\}$	4.0	$(0.57, 7.40)$
3	$\{0, 0, 0, -6, -6, -0.5, -0.5\}$	3.99	$(0.41, 7.57)$
4	$\{-6.6, 0, 4.5, 5.2, 5.2, 0.14, 0.14\}$	3.91	$(0.46, 7.36)$

Table 3.5: Analogous to Table 3.3, types of sensitivity parameters that link distributions of a binary outcome y for respondents, $f(y \mid \boldsymbol{x}, d = 0; \boldsymbol{\theta})$, and nonrespondents, $f(y \mid \boldsymbol{x}, d = 1; \boldsymbol{\theta})$.

Type	Modification	Description
1	$\text{logit}(P\{y = 1 \mid d = 1, \boldsymbol{\theta}\}) = \text{logit}(P\{y = 1 \mid d = 0, \boldsymbol{\theta}\}) + \delta$	Marginal odds scaling
2	$\text{logit}(P\{y = 1 \mid \boldsymbol{x}, d = 1, \boldsymbol{\theta}\}) = \text{logit}(P\{y = 1 \mid \boldsymbol{x}, d = 0, \boldsymbol{\theta}\}) + \delta_{\chi}$	Conditional odds scaling
3	$\text{logit}(P\{y = 1 \mid d = 1, \boldsymbol{\theta}\}) = \text{logit}(P\{y = 1 \mid d = 0, \boldsymbol{\theta}\}) + \beta_{x_j} x_j$	Marginal effect change for x_j
4	$\text{logit}(P\{y = 1 \mid \boldsymbol{x}, d = 1, \boldsymbol{\theta}\}) = \text{logit}(P\{y = 1 \mid \boldsymbol{x}, d = 0, \boldsymbol{\theta}\}) + \beta_{\chi, x_j} x_j$	Conditional effect change for x_j

have firm scientific grounding, then the conclusion of no effect under MCAR and MAR models should be carefully reviewed and, possibly, declared unreliable. This result highlights the importance of sensitivity analyses, especially for studies with a substantial fractions of missing outcomes and borderline conclusions.

Similar ideas can be applied to binary outcomes with slight differences in the interpretation of sensitivity parameters. Table 3.5 defines some sensitivity parameters, analogous to the ones in Table 3.3, for a binary outcome modeled using logistic link-function. All sensitivity parameters affect the odds of success for nonrespondents. For example, type 1 may be used if the odds of success for nonrespondents are believed

to differ by a factor of e^{δ} from the odds for respondents. The exploration of alternative models for binary outcomes can be done in the same manner as described for continuous outcomes; see Section 2.3.2 for a real-data example with type 2 sensitivity parameters.

3.4 Software for ETP Displays

There are several R-packages available to produce MIs under the MCAR or MAR models[1]. Some of them use fully conditional specifications: `mice, mi, BaBooN`. Others attempt to model the data jointly: `amelia, MImix, mix, norm, pan`. R-package `SensMice` contains a function `sens.mice` (Resseguier et al. 2011) which allows users to modify the imputation model, after it was estimated automatically under the MAR assumption, for the purpose of performing sensitivity analysis. Available modifications are analogous to type 1 in Tables 3.3 and 3.5.

Among stand-alone software packages that perform multiple imputation are SOLAS 4.0, IVEware and WinMICE, which use chained equations, S-PLUS and NORM, which model the data jointly. However, none of the existing versions of these packages offers any type of systematic sensitivity analyses. We have implemented an R-procedure that draws ETP displays, as illustrated in this chapter, for generated MIs. The procedure is available for download from `http://sites.google.com/site/vliublinska/research`. Statistical Solutions Ltd are planning to implement the sensitivity analysis capability, based on the ETP displays and the procedure described in Section 3.3, into the SOLAS package.

[1]See `http://www.stefvanbuuren.nl/mi/Software.html` for a complete and up-to-date list

3.5 Discussion

In this chapter we generalized a method of performing systematic sensitivity analyses using ETP displays for studies with partially missing outcomes. As before, the method requires a dichotomous treatment and a set of fully-observed predictors. We showed that there is a substantial flexibility in the types of distributions of the outcome that are suitable for the ETP displays. In addition to the family of one-parameter NEF distributions, as well as any of their one-to-one transformations, we demonstrated how the displays can be adapted to use with a two-parameter Normal distribution of the outcome.

We also described an intuitive way to explore MNAR models by utilizing the pattern-mixture factorization of the posterior distribution of the outcomes. Under the MAR assumption, the distributions of the outcomes for respondents and nonrespondents, conditional on the same set of covariates, coincide, and the former can be used as a baseline to construct many alternative models for the latter, by introducing various sensitivity parameters. In fact, ETP displays themselves may suggest possible directions of high sensitivity for building alternative models that will change the study's conclusions.

There are many software packages that produce MIs for a data set with missing values, e.g. SOLAS, IVEWare, MICE. We developed an R-procedure that draws ETP displays using prespecified imputations. To summarize, the proposed displays help reveal the weakness, or confirm the strength, of the conclusions of the study under the adopted assumptions and guide the consideration of alternative models that can alter the conclusions.

Chapter 4

Principal Stratification as a Method of Sensitivity Analysis in Studies with Missing Data

4.1 Introduction

In Chapters 2 and 3 we introduced enhanced tipping-point displays to help systematize the exploration of alternative models for data with missing values to assess the strength of the drawn conclusions. Here we continue exploring methods of sensitivity analyses and demonstrate the use of principal stratification framework for this purpose.

Principal stratification (PS), first described by Frangakis and Rubin (2002), is a general framework of adjusting the estimation of causal estimands based on post-treatment outcomes. In its simplest form, this framework builds on the Rubin Causal

Model (RCM, Holland 1986) by explicitly identifying latent classes (i.e., *principal strata*) of units based on a categorization of all posttreatment outcomes. Potential outcomes are then modeled separately for each principal stratum. Current applications of PS exist for several types of posttreatment outcomes, e.g., treatment noncompliance (Imbens and Angrist 1994; Imbens and Rubin 1997), missingness in outcomes (Jin and Rubin 2009), and censoring "due to death" (Zhang and Rubin 2003; Rubin 2006a; Zhang et al. 2008, 2009). All three of the complications above were recently considered simultaneously in a study of causal effects of a job-training program on employment and wages by Frumento et al. (2012)

Here we demonstrate a novel application of PS as a method of sensitivity analysis. Several examples of the use of PS in the context of sensitivity analysis has been given in the literature. Majority of them deal with one complication at a time, e.g., noncompliance (Egleston et al. 2010), other intermediate outcome (Gilbert et al. 2003; Hudgens et al. 2003; Shepherd et al. 2008), or censoring (Shepherd et al. 2007), and handle a basic no-covariate settings. We also found an R-package `sensitivityPStrat` that provides methods to perform sensitivity analyses of treatment effects within principal strata described in some references mentioned above. The application presented here handles two complications at once, which result in many more strata than in existing examples. More importantly, our example incorporates covariates into the outcome models, which substantially complicates the inference and requires an improved method of model fitting introduced below.

Our application of PS is demonstrated on a clinical trial, described in Section 2.3.2, that had missingness in outcomes due to death. Recall that initial analysis of

this trial was conducted under the assumption of no distinction between potential outcomes missing due to death and outcomes missing due to other reasons (e.g., lost to follow-up or missed appointments). However, as discussed by Rubin (1998, Section 6), Frangakis and Rubin (2002) and Zhang and Rubin (2003), such an assumption is inappropriate, because potential outcomes for deceased subjects are not well-defined.

Undefined potential outcomes can arise in other settings, besides subject death during the course of a study. For example, subject wages if unemployed, a miscarriage for a women that is not pregnant, or a college graduate-point average (GPA) for high-school drop-outs that did not get a General Equivalency Diploma are also undefined. In these examples, "survival status" of a study subject (employment status, pregnancy status, or drop-out indicator, correspondingly) is a posttreatment outcome, and is crucial for defining and modeling potential outcomes of ultimate interest. As such, we apply PS framework in our analysis to address the censoring due to death, and to assess sensitivity of the study's conclusions to alternative potential outcome model specifications.

A second issue addressed in our application is the inherent difficulty of posterior computations under PS. The current PS literature contains many analyses that emphasize the complex computations required for estimation of causal estimands (e.g., Barnard et al. 2003; Jin and Rubin 2009; Gallop et al. 2009; Zhang et al. 2009; Elliott et al. 2010; Frumento et al. 2012). As a result of the fundamental problem of causal inference, latent principal strata can never be fully observed, and can only be inferred from background covariates and observed outcomes. Consequently, a model-based PS analysis can involve weakly identified models, and the data may contain little infor-

mation that helps characterize principal strata, or the sample size may not even be sufficient for estimation purposes relative to the number of strata that are formed; all these issues will slow down posterior computations.

The novelty of our approach lies in using Hamiltonian Monte Carlo algorithm, originally called a Hybrid Monte Carlo (HMC, Duane et al. 1987), to obtain draws from posterior distributions of interest. This algorithm can be viewed as a data-augmentation (DA) strategy, because a vector of parameters, considered as a "position" variable, is augmented with fully missing "momentum" vector, and Hamiltonian dynamics are then used to perform a more effective exploration of the posterior. When the gradient of the logarithm of the posterior exists, HMC can reduce the correlation between successive draws considerably (Neal 1995) and converge faster than a commonly used Metropolis-Hastings Monte Carlo (MHMC) method, even when the support of the posterior has substantial curvature (Neal 2011). We demonstrate the superiority of HMC over MHMC for posterior computations under the PS using data from two real examples. Then we apply the HMC algorithm to the real data collected from the medical device clinical trial and estimate various estimands of interest.

The rest of this chapter is organized as follows. In Section 4.2 we briefly summarize the initial analyses performed for the clinical trial under consideration and describe the issues not considered originally, especially the missingness due to death. In Section 4.3 we define principal strata, specify models for potential outcomes, and list the attendant assumptions. Section 4.4 describes our HMC algorithm for PS, and includes a comparison of the performance of HMC to that of a standard MHMC for two examples, with further details provided in Appendices C.1, C.2 and C.3. Section 4.5

describes results obtained for the clinical trial under consideration using HMC and concludes with a discussion.

4.2 Description of the Clinical Trial

Our primary data comes from a noninferiority clinical trial described in Section 2.3.2. As it was noted, analysis of the primary outcome of interest, i.e. the frequency of cement leakage, indicated that the new treatment resulted in significantly fewer leaks. Then we focused on the secondary outcomes of interest, i.e., six adverse events, posttreatment pain scores and disability levels, that had large fractions of missing data, primarily as a result of missed follow-up appointments or patient death. The initial analysis of secondary endpoints was performed under the MAR assumption, ignoring the issue of censoring due to death. The resulting conclusion was that the data provides no evidence for differences in rates of adverse events, average posttreatment pain scores, or disability levels between the treatment and control groups.

We also demonstrated that there is essentially no sensitivity of this conclusion to various deviations from the initial MAR assumption. We now perform a refined sensitivity analysis, recognizing the fact that missingness due to death requires a fundamentally different consideration. In the next section we provide a detailed description of the application of PS to address this issue.

4.3 Application of Principal Stratification to the Clinical Trial

4.3.1 Notation and Identification of Principal Strata

The clinical trial under consideration had $N = 77$ subjects, including $N_T = 49$ subjects that received the new treatment and $N_C = 28$ that underwent a previously approved procedure. Here we denote the vector of treatment assignment indicators by $\mathbf{Z} = (z_1, z_2, \ldots, z_N)'$, such that

$$z_i = \begin{cases} 1 & \text{if subject } i \text{ received the new treatment,} \\ 0 & \text{otherwise,} \end{cases}$$

$i = 1, \ldots, N$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})'$ be the vector of all pretreatment covariates for subject i, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ be a $K \times N$ matrix of covariates for all subjects.

Several outcomes of interest were collected for each subject at two time points after the surgery, three and twelve months. The first set of potential outcomes that we consider is censoring due to death. For subject i at time point $t \in \{1, 2\}$ under treatment $z \in \{0, 1\}$ a potential outcome $d_{i,t}(z)$ is defined as

$$d_{i,t}(z) = \begin{cases} 1 & \text{if subject } i \text{ is deceased at time } t \text{ under treatment } z, \\ 0 & \text{otherwise,} \end{cases}$$

The value observed in the study is then

$$d_{i,t} = d_{i,t}(1)z_i + d_{i,t}(0)(1 - z_i),$$

Table 4.1: Principal strata generated by the censoring due to death. Of the total number $2^4 = 16$ of possible combinations, only nine distinct strata arise, because death at time $t = 1$ automatically implies death at $t = 2$.

k	$d_{i,1}(1)$	$d_{i,2}(1)$	$d_{i,1}(0)$	$d_{i,2}(0)$
1	0	0	0	0
2	0	0	0	1
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	1
7	1	1	0	0
8	1	1	0	1
9	1	1	1	1

a function of the potential outcomes and observed treatment assignment. We let $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ be the observed vectors of death indicators, $\boldsymbol{D}_t = (d_{1,t}, d_{2,t}, \ldots, d_{N,t})'$.

Principal stratum for subject i are defined by the vector

$$s_i = (d_{i,1}(1), d_{i,2}(1), d_{i,1}(0), d_{i,2}(0))'.$$

For example, if subject i would be alive at 12 months after receiving a new treatment, but would be deceased at 3 months follow-up if administered a control procedure, then $s_i = (0, 0, 1, 1)'$. Alternatively, knowing the principal stratum of subject i immediately determines the survival status under any treatment. Table 4.1 shows nine possible strata that arise from different combinations of the survival status.

The second set of potential outcomes constitutes the main focus of our analysis. For subject i at time point t under treatment z we define a vector of secondary potential outcomes

$$\boldsymbol{y}_{i,t}(z) = (y_{1i,t}(z), y_{2i,t}(z), y_{3i,t}(z))',$$

where $y_{1i,t}(z)$ represents a number of adverse events out of at total of five considered, $y_{2i,t}(z)$ is the pain score (a number between 0 and 10, with 0 being no pain), and $y_{3i,t}(z)$ is the disability index (a number between 0 and 100, with 0 indicating no disability). We let $\boldsymbol{Y}_t(z)$ be a $3 \times N$ matrix of secondary potential outcomes for all subjects, $\boldsymbol{Y}_t(z) = (\boldsymbol{y}_{1,t}(z), \boldsymbol{y}_{2,t}(z), \ldots, \boldsymbol{y}_{N,t}(z))$. Note that if $d_{i,t}(z) = 1$, then $y_{1i,t}(z)$, $y_{2i,t}(z)$, $y_{3i,t}(z)$ are all undefined.

Finally, we consider missingness in outcomes not due to death and define potential outcomes that indicate missingness due to other reasons. Let

$$m_{i,t}(z) = \begin{cases} 1 & \text{if all components of } \boldsymbol{y}_{i,t}(z) \text{ are well-defined but missing,} \\ 0 & \text{otherwise,} \end{cases}$$

and let a vector $\boldsymbol{M}_t(z) = (m_{1,t}(z), m_{2,t}(z), \ldots, m_{N,t}(z))'$ contain missingness indicators at time t under treatment z for all subjects in the study. Analogous to principal strata that arise from the survival status, further principal stratification can be introduced on the basis of potential missingness, with strata defined as $(m_{i,1}(1), m_{i,2}(1), m_{i,1}(0), m_{i,2}(0))'$. A total of 16 strata are generated by this approach, as summarized in Table 4.2.

However, such level of generalization introduces considerable challenges in the estimation of outcome models for each individual stratum, because it asserts that some outcomes will never be observed, e.g., outcomes for subjects that belong to strata 15 and 16 in Table 4.2. Therefore, their models can not be estimated without substantial unassessable assumptions. Moreover, the total number of possible strata, defined both by survival status and potential missingness, is $9 \cdot 16 = 144$. Although this is

Table 4.2: Principal strata generated by potential missingness in potential outcomes due to other reasons besides death, observed at two time-points under the dichotomous treatment.

j	$m_{i,1}(1)$	$m_{i,2}(1)$	$m_{i,1}(0)$	$m_{i,2}(0)$
1	0	0	0	0
2	0	0	0	1
3	0	0	1	0
...	...	...	...	...
15	1	1	1	0
16	1	1	1	1

theoretically the most complete specification of principal strata in the clinical trial under consideration, the sample size of only $N = 77$ would make the analysis practically infeasible. Therefore, we do not consider the generalization due to potential missingness and proceed with the analysis under the original MAR assumptions.

Table 4.3 summarizes our notation and provides an example of a "Science" (Rubin 2007) of the study,

$$(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{D}_1(T), \boldsymbol{D}_2(T), \boldsymbol{D}_1(C), \boldsymbol{D}_2(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(T), \boldsymbol{Y}_1(C), \boldsymbol{Y}_2(C)),$$

a collection of all pretreatment covariates, treatment and stratum indicators, and potential outcomes, for a particular realization of the treatment assignment. Next, we list assumptions utilized in our analysis.

4.3.2 Assumptions and Estimands of Interest

The following assumptions are necessary for using PS framework:

- *Stable Unit Treatment Value Assumption* (SUTVA, Rubin 1980). Potential

Table 4.3: The example of the Science for subjects in the study for a particular realization of the treatment assignment. Note that $d_{i,t}(z)$ and $\boldsymbol{y}_{i,t}(z)$ are always missing simultaneously.

i	$\boldsymbol{x}_i$	s_i	z_i	$d_{i,1}(1)$	$d_{i,2}(1)$	$d_{i,1}(0)$	$d_{i,2}(0)$	$\boldsymbol{y}_{i,1}(1)$	$\boldsymbol{y}_{i,2}(1)$	$\boldsymbol{y}_{i,1}(0)$	$\boldsymbol{y}_{i,2}(0)$
1	*	?	1	*	*	?	?	*	*	?	?
2	*	?	1	*	*	?	?	*	*	?	?
...	*	?	1	*	??	?	?	*	??	?	?
N_T	*	?	1	*	*	?	?	*	*	?	?
N_T+1	*	?	0	?	?	??	??	?	?	??	??
N_T+2	*	?	0	?	?	*	*	?	?	*	*
...	*	?	0	?	?	??	*	?	?	??	*
N_T+N_C	*	?	0	?	?	*	*	?	?	*	*

* indicates observed values, ? indicates unobserved values, ?? indicates missing values not due to death.

outcomes of any specific subject do not depend on other subjects' treatment assignments, i.e., for any two vectors of treatment assignments $\boldsymbol{Z}, \boldsymbol{Z'} \in \{0,1\}^N$, with $z_i = z'_i$, SUTVA states that

$$\boldsymbol{y}_{i,t}(\boldsymbol{Z'}) = \boldsymbol{y}_{i,t}(\boldsymbol{Z}) \text{ and } d_{i,t}(\boldsymbol{Z'}) = d_{i,t}(\boldsymbol{Z}), \text{ for } t = 1, 2,$$

where $y_{i,t}(\boldsymbol{Z})$ are $d_{i,t}(\boldsymbol{Z})$ are vectors of potential outcomes for unit i under treatment assignment $\boldsymbol{Z}$ for all units in the study. SUTVA also requires that there is only one well-defined version of each treatment (e.g., no dose variations). Both conditions ensure that there are only two potential outcomes for each subject at each time point, corresponding to the two possible treatments. There are many plausible scenarios that would result in a violation of this assumption in the trial. For example, there is no information about doctors that performed the surgeries: if some doctors treated more than one patient, their technique could have improved with time, and so potential outcomes of later patients

would effectively have depended on the treatment assigned to subjects entering the study earlier. Given lack of evidence that contradicts SUTVA, we make this assumption in our analysis.

- *Unconfoundness* (Rubin 1990). Treatment assignment depends only on observed pretreatment covariates:

$$\boldsymbol{Z} \mid \boldsymbol{D}_1(0), \boldsymbol{D}_2(0), \boldsymbol{D}_1(1), \boldsymbol{D}_2(1), \boldsymbol{Y}_1(0), \boldsymbol{Y}_2(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(1), \boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{X} \sim \boldsymbol{Z} \mid \boldsymbol{X}.$$

Randomization performed in the trial justifies this assumption. By design, the distribution of the treatment indicators $\boldsymbol{Z}$ is

$$f(\boldsymbol{Z} \mid \boldsymbol{X}) = f(\boldsymbol{Z}) = \begin{cases} 1/\binom{N}{N_T} & \text{if } \sum z_i = N_T, \\ 0 & \text{otherwise.} \end{cases}$$

In fact, because $0 < P(z_i = 1) < 1$ for all i, the assignment mechanism is also *strongly ignorable* (Rosenbaum and Rubin 1983).

- *Ignorable missingness mechanism* (Rubin 1976; Little and Rubin 2002). Missingness in outcomes is MAR, and parameters that govern the missingness and potential outcomes models are distinct (see Section 1.1). Let $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ be vectors of missingness indicators at three and twelve months after the surgery, respectively, and let

$$f(\boldsymbol{M}_1, \boldsymbol{M}_2 \mid \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(1), \boldsymbol{Y}_1(0), \boldsymbol{Y}_2(0), \boldsymbol{D}_1(1), \boldsymbol{D}_2(1), \boldsymbol{D}_1(0), \boldsymbol{D}_2(0), \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\phi}) \tag{4.1}$$

be the conditional distribution of the missingness indicators, governed by the vector-parameter $\boldsymbol{\phi}$, given the rest of the Science. Formally, ignorability implies that (4.1) equals

$$f(\boldsymbol{M}_1, \boldsymbol{M}_2 \mid \boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{obs}, \boldsymbol{D}_1^{obs}, \boldsymbol{D}_2^{obs}, \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\phi}),$$

where $\boldsymbol{Y}_1^{obs}$, $\boldsymbol{Y}_2^{obs}$, $\boldsymbol{D}_1^{obs}$, $\boldsymbol{D}_2^{obs}$ are the observed outcome values and death indicators at each time-point, respectively. Ignorability is unassessable, i.e., the observed data themselves cannot confirm or contradict it without additional assumptions. Ignorability is assumed in the trial under consideration to simplify the analysis, however, this assumption can be relaxed by assuming *latent ignorability* (Frangakis and Rubin 1999), where missingness is related to death. Although latent ignorability may be more realistic, it would complicate computation and modeling considerably.

In Section 4.3.1 we identified nine principal strata that arise from the survival status. We reduce the number of strata by making the following assumption.

- *Monotonicity of the treatment effect on death* (Zhang and Rubin 2003). Assignment to the new treatment results in the same or better survival status than assignment to the control, i.e. $d_{i,t}(1) \geq d_{i,t}(0)$, for $t = 1, 2$ and $i = 1, \ldots, N$. As shown in Table 4.4, this assumption reduces the number of strata from nine to six. Under monotonicity, we redefine latent stratum indicators s_i as each taking a value from the following set $\Omega = \{aa, pa, pp, na, np, nn\}$.

Table 4.4: Final set of strata that arise from having some outcomes censored due to death and assuming monotonicity. Here, "a" stands for always-survivor, "p" for partial survivor and "n" for never-survivor, and the position of the letter corresponds to the time period.

k	$d_{i,1}(0)$	$d_{i,2}(0)$	$d_{i,1}(1)$	$d_{i,2}(1)$	Stratum label
1	0	0	0	0	aa
2	0	1	0	0	pa
3	1	1	0	0	na
4	0	1	0	1	pp
5	1	1	0	1	np
6	1	1	1	1	nn

As discussed earlier, it is important to define latent strata based on survival status because outcomes for diseased subjects are not well-defined. This fact limits meaningful estimands to particular principal strata. Here, we focus on the following finite-population estimands:

- Average treatment effects on the number of adverse events, pain score, and disability index at 3 months for partial and always survivors,

$$\boldsymbol{\delta}_1 = \sum_{i:\ s_i \in \{aa,pa,pp\}}^{N} \{\boldsymbol{y}_{i,1}(1) - \boldsymbol{y}_{i,1}(0)\} / \sum_{i:\ s_i \in \{aa,pa,pp\}}^{N} 1.$$

  Each component of this vector of estimands can also be viewed as a weighted average of separate treatment effects for the three strata aa, pa, and pp.

- Average treatment effects on the number of adverse events, pain score, and disability index at 12 months for always-survivors in both treatment groups,

$$\boldsymbol{\delta}_2 = \sum_{i:\ s_i = aa}^{N} \{\boldsymbol{y}_{i,2}(1) - \boldsymbol{y}_{i,2}(0)\} / \sum_{i:\ s_i = aa}^{N} 1.$$

In addition, we are interested in the following descriptive estimands:

- Probability of death within 3 months under the treatment and under the control,

$$\xi_1^z = \sum_{i:z_i=z}^{N} d_{i,1}(z) / \sum_{i:z_i=z}^{N} 1 = \begin{cases} \sum_{i=1}^{N} I\left(s_i \in \{na, np, nn\}\right)/N & \text{if } z = 0, \\ \sum_{i=1}^{N} I\left(s_i = nn\right)/N & \text{if } z = 1. \end{cases}$$

- Chance of death between 3 and 12 months under the treatment and under the control,

$$\xi_2^z = \sum_{i:z_i=z}^{N} \{d_{i,2}(z) - d_{i,1}(z)\} / \sum_{i:z_i=z}^{N} 1 = \begin{cases} \sum_{i=1}^{N} I\left(s_i \in \{pa, pp\}\right)/N & \text{if } z = 0, \\ \sum_{i=1}^{N} I\left(s_i \in \{np, pp\}\right)/N & \text{if } z = 1. \end{cases}$$

4.3.3 Model Specifications for Potential Outcomes and Principal Strata Membership

Principal strata are usually only partially observed, however, we can identify a set of strata that correspond to each combination of observed survival statuses for subjects in the treatment or the control groups. Table 4.5 groups the observed data in our study and lists the corresponding latent strata for each group.

We proceed with specifying a Bayesian model for the potential outcomes, separately for each principal stratum. A full joint distribution of the potential outcomes

Table 4.5: Observed subject groups, corresponding principal strata and the number of subjects in each group. Here, ?? indicates missing values that could have been observed.

Treatment group	Observed outcomes groups $O(d_{i,1}, d_{i,2})$	Possible latent strata	Number of subjects ($N = 77$)
T	$O(0,0)$	aa, pa, na	25
T	$O(0,1)$	np, pp	3
T	$O(1,1)$	nn	0
C	$O(0,0)$	aa	17
C	$O(0,1)$	pp, pa	1
C	$O(1,1)$	na, np, nn	2
T	$O(0,??)$	aa, pa, pp, na, np	9
T	$O(1,??)$	nn	0
T	$O(??,0)$	aa, pa, na	3
T	$O(??,1)$	np, pp, nn	2
T	$O(??,??)$	aa, pa, pp, na, np, nn	7
C	$O(0,??)$	aa, pp, pa	5
C	$O(1,??)$	aa, np, nn	0
C	$O(??,0)$	aa	2
C	$O(??,1)$	pa, pp, na, np, nn	0
C	$O(??,??)$	aa, pa, pp, na, np, nn	1

can be partitioned as follows,

$$f(\boldsymbol{Y}_1(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(C), \boldsymbol{Y}_2(T), \boldsymbol{D}_1(C), \boldsymbol{D}_1(T), \boldsymbol{D}_2(C), \boldsymbol{D}_2(T), \boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{Z}, \boldsymbol{S} \mid \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\phi}) =$$

$$f(\boldsymbol{Y}_1(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(C), \boldsymbol{Y}_2(T), \boldsymbol{Z}, \boldsymbol{S} \mid \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\phi}) \cdot$$

$$f(\boldsymbol{M}_1, \boldsymbol{M}_2 \mid \boldsymbol{Y}_1(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(C), \boldsymbol{Y}_2(T), \boldsymbol{D}_1(C), \boldsymbol{D}_1(T), \boldsymbol{D}_2(C), \boldsymbol{D}_2(T), \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\phi}) =$$

$$f(\boldsymbol{Z}) f(\boldsymbol{Y}_1(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(C), \boldsymbol{Y}_2(T), \boldsymbol{S} \mid \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}) \cdot$$

$$f(\boldsymbol{M}_1, \boldsymbol{M}_2 \mid \boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{obs}, \boldsymbol{D}_1^{obs}, \boldsymbol{D}_2^{obs}, \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\phi})$$

Last equality holds by the ignorability assumption and randomization. Because we are interested in estimating $\boldsymbol{\theta}$ only, we can drop the model for the missingness mechanism

and consider the following part only,

$$f(\boldsymbol{Y}_1(C), \boldsymbol{Y}_1(T), \boldsymbol{Y}_2(C), \boldsymbol{Y}_2(T), \boldsymbol{S} \mid \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}). \tag{4.2}$$

Principal stratum membership is modeled using multinomial distribution with logistic link-function and four predictors $x_{i1}, \ldots, x_{i4}$, corresponding to age, sex, the interaction of age and sex, and BMI,

$$\Pr(s_i = s \mid \boldsymbol{x}_i; \boldsymbol{\psi}) = \frac{\exp(\psi_{s,0} + \psi_{s,1}x_{i1} + \psi_{s,2}x_{i2} + \psi_{s,3}x_{i3} + \psi_{s,4}x_{i4})}{\sum_{h\in\Omega}\exp(\psi_{h,0} + \psi_{h,1}x_{i1} + \psi_{h,2}x_{i2} + \psi_{h,3}x_{i3} + \psi_{h,4}x_{i4})}, \tag{4.3}$$

where $s \in \Omega$ and $\boldsymbol{\psi}$ is a vector of all parameters. We let $\psi_{nn,0} = \psi_{nn,1} = \cdots = \psi_{nn,4} = 0$, so that stratum nn is taken as a baseline. Model (4.3) requires estimating $5 \cdot 5 = 25$ parameters and, in order to simplify it, we assume that the slopes that correspond to each predictor are identical across principal strata: $\psi_k \equiv \psi_{s,k} = \psi_{\tilde{s},k}$ for all $s, \tilde{s} \in \Omega$ and for $k = 1, 2, 3, 4$. This restriction reduces the number of parameters to $4 + 5 = 9$, making the subsequent estimation more feasible.

Next, we introduce the model for the vector of potential outcomes $\boldsymbol{y}_{i,t}(z) = (y_{1i,t}(z), y_{2i,t}(z), y_{3i,t}(z))'$, $z \in \{0, 1\}$ and $t = 1, 2$. Here, $y_{1i,t}(z)$ represents a number (0-5) of adverse events observed for unit i within 3 months ($t = 1$) or between 3 and 12 months ($t = 2$) after the surgery, if assigned to group z. Therefore, we use Binomial distribution to model these outcomes, conditional on stratum s_i:

$$\begin{aligned} y_{1i,t}(z) \mid s_i, \boldsymbol{x}_i, \boldsymbol{\beta} &\sim \text{Binom}(q_{i,t}(z), 5), \\ \text{logit}(q_{i,t}(z)) &= \beta_{s_i,1tz,0} + \beta_{s_i,1tz,1}x_{i5} + \beta_{s_i,1tz,2}x_{i6}, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\beta}$ is a vector of all parameters, and $x_{i,5}$, $x_{i,6}$ are baseline pain score and disability index.

Last two outcomes of interest, $y_{2i,t}(z)$ and $y_{3i,t}(z)$, represent pain scores and disability indexes. In order to be eligible for the study, subjects were required to have certain minimum pretreatment pain score and disability index, and a desirable outcome of the surgery was to reduce both characteristics to zero. Both treatment and control procedures in the study were very effective in reducing pain and disability and, as a result, the distributions of these measures have a large point mass at zero (see Figure 4.1).

For simplicity, we consider new outcomes that indicate whether pain and disability were fully eliminated, $\tilde{y}_{ri,t}(z) = I(y_{ri,t}(z) = 0)$, and model them as follows:

$$\begin{aligned} \tilde{y}_{ri,t}(z) \mid s_i, \boldsymbol{x_i}, \boldsymbol{\beta} &\sim \text{Bern}(u_{ri,t}(z)), \\ \text{logit}(u_{ri,t}(z)) &= \beta_{s_i,rtz,0} + \beta_{s_i,rtz,1} x_{i5} + \beta_{s_i,rtz,2} x_{i6}. \end{aligned} \tag{4.5}$$

Note that all three outcomes, $y_{ri,t}(z)$ with $r = 1, 2, 3$, are defined in the following cases only:

- $t = 1$, $z = 0$ and subject i is in stratum $s_i \in \{aa, pa, pp\}$,
- $t = 1$, $z = 1$ and subject i is in stratum $s_i \in \{aa, pa, na, pp, np\}$,
- $t = 2$, $z = 0$ and subject i is in stratum $s_i = aa$,
- $t = 2$, $z = 1$ and subject i is in stratum $s_i \in \{aa, pa, na\}$.

Considering that $\tilde{y}_{ri,t}(z)$ can be correlated across strata and time, we use slopes to

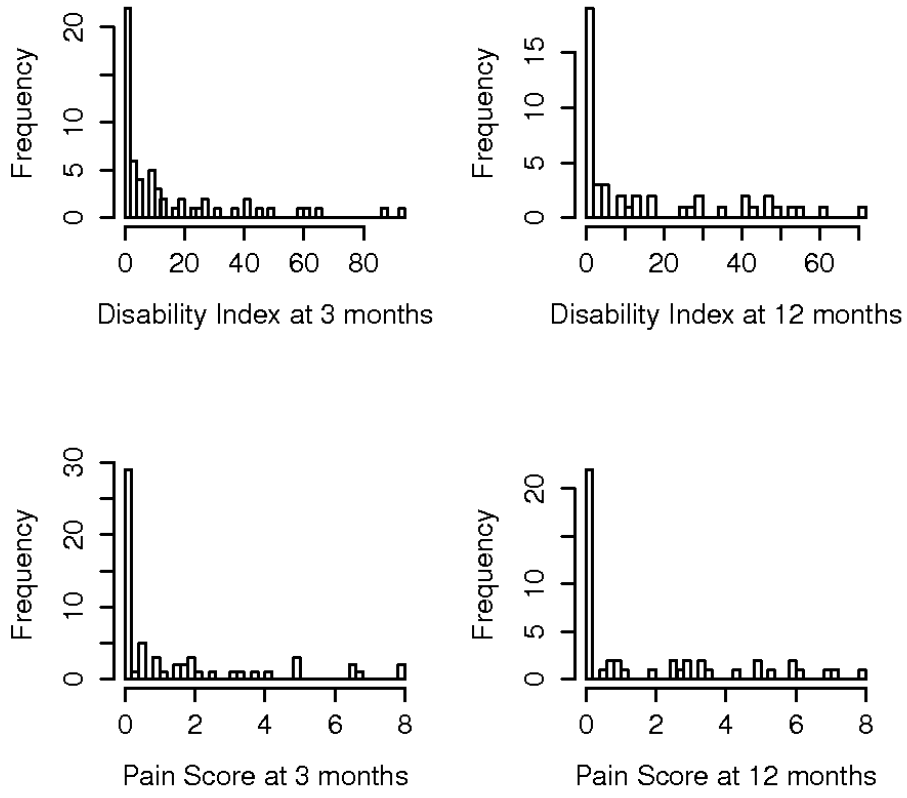


Figure 4.1: Histograms of pain scores and disability indexes, recorded at 3 and 12 months after the surgery. Note the point mass at zero for all four outcomes.

reflect this fact, by assuming that $\beta_k \equiv \beta_{s,rtz,k} = \beta_{\tilde{s},\tilde{r}\tilde{t}\tilde{z},k}$ for all $r, \tilde{r} = 1, 2, 3$ and $k = 1, 2, 3$, and combinations of $(s, t, z), (\tilde{s}, \tilde{t}, \tilde{z}) \in \Omega \times \{1, 2\} \times \{0, 1\}$ for which the outcomes exist. In other words, the slopes for each predictor are the same across all potential outcomes. This assumptions reduced the number of parameters for models (4.4) and (4.5) from $3 \times 3 \times 12 = 108$ to $3 \times 12 + 2 = 38$. To summarize, the total number of parameters required to model (4.2) is 47 (see Appendix C.1). In the next section we introduce HMC, the computational method used to perform model-fitting and compare its performance to that of a standard MHMC.

4.4 Application of HMC Method to PS Computations

4.4.1 General Overview

HMC method uses Markov Chain Monte Carlo (MCMC) technique to efficiently sample from complex joint distributions with highly-correlated parameters using Gibbs sampling, combined with the acceptance rule from MHMC method and ideas from Hamiltonian dynamics. Suppose the goal is to sample $\boldsymbol{\theta} \in \Re^J$ from $\pi(\boldsymbol{\theta}) \propto \exp\{-U(\boldsymbol{\theta})\}$. The HMC method considers an artificial dynamic system, viewing $\boldsymbol{\theta}$ as position coordinates of a particle with potential pseudo-energy equal to $U(\boldsymbol{\theta})$. In addition, it introduces auxiliary momentum vector $\boldsymbol{p} \in \Re^J$ and defines a kinetic pseudo-energy of the particle as $k(\boldsymbol{p}) = \boldsymbol{p}'\Lambda^{-1}\boldsymbol{p}/2$, where Λ is a $J \times J$ positive-definite "mass matrix" (e.g., if it is diagonal, its elements can be viewed as "masses" of each component of $\boldsymbol{\theta}$).

MCMC sampling is performed on the augmented parameters space $(\boldsymbol{\theta}, \boldsymbol{p})$ that has the following distribution,

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}) \propto \exp\{-H(\boldsymbol{\theta}, \boldsymbol{p})\},$$

where $H(\boldsymbol{\theta}, \boldsymbol{p}) = U(\boldsymbol{\theta}) + k(\boldsymbol{p})$. The function $H(\boldsymbol{\theta}, \boldsymbol{p})$ is called *Hamiltonian*, it represents the total energy of the particle. Marginally, it can be shown that $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{p} \sim \text{Norm}_J(\mathbf{0}, \Lambda)$. Sampling rules are derived from the law of the conservation of energy, that says that the total energy remains constant in a closed system. The

advantage of this method is that the resulting MCMC moves follow the dynamics of the target distribution more closely: produced proposal are distant and have high probability of acceptance, which results in more efficient exploration of the target distribution (Liu 2008, Ch. 9).

An important aspect of our HMC implementation is the choice of the mass matrix Λ. As discussed in Girolami and Calderhead (2011), the efficiency of the algorithm can be increased if the mass matrix reflects the curvature of the target distribution. At the beginning of the Gibbs step that generates a draw from the conditional posterior of $\boldsymbol{\theta}$, we first find the minimum of the negative logarithm of the posterior and evaluate the Hessian of this function at the minimum. This choice of mass matrix for HMC algorithm results in a converging Markov chain, with a stationary distribution that is the desired posterior $\pi(\boldsymbol{\theta})$, as justified by Theorem 1 in Burda and Maheu (2011).

As noted in Section 4.1, posterior computations under PS can be difficult to implement, because principal strata are not necessarily fully observed for all subjects in a study and have to be inferred from background covariates and observed outcomes. Next we demonstrate the application of the HMC-within-Gibbs algorithm to calculating posterior distribution of estimands of interest with two examples that have simpler data structures (two and three principal strata defined by non-compliance).We also use these examples to compare the performance of our algorithm to that of a standard MHMC algorithm.

4.4.2 Example 1: Canvassing and Voter Turnout

For our first example we use a randomized factorial experiment performed by Gerber and Green (2000, 2005) to study the effects of non-partisan mail, canvassing, and phone calls on voter turnout. For illustration purposes, we focus on canvassing only, and compare turnout behavior for subjects assigned to the canvassing treatment against those assigned to no treatment. Further details of this analysis are given in Gill et al. (2013).

The data consists of $N = 6,617$ experimental subjects randomly assigned to canvassing ($z_i = 1$) or not ($z_i = 0$). A subject assigned to be canvassed can refuse to comply. Let $d_i(z)$ indicate if subject i is actually canvassed under treatment z. If $d_i(z) = z$, then a subject is said to be a complier, whereas if $d_i(0) = d_i(1) = 0$, then a subject is said to be a never-taker. By design of the experiment, it is impossible for a subject to be canvassed when assigned control, i.e., for $d_i(0) = 1$. Therefore, only two principal strata are formed: compliers and never-takers, represented by $s_i = c$ if $(d_i(0), d_i(1))' = (0, 1)'$ and $s_i = n$ if $(d_i(0), d_i(1))' = (0, 0)'$, respectively.

Potential outcome $y_i(z)$ is defined as

$$y_i(z) = \begin{cases} 1 & \text{if subject } i \text{ voted in the election under treatment } z, \\ 0 & \text{otherwise.} \end{cases}$$

Estimands of interest are finite population average causal effects of canvassing for never-takers and compliers, defined as

$$\sum_{i:s_i=n} \{y_i(1) - y_i(0)\} \Big/ \sum_{i:s_i=n} 1 \text{ and } \sum_{i:s_i=c} \{y_i(1) - y_i(0)\} \Big/ \sum_{i:s_i=c} 1. \tag{4.6}$$

Note that compliance status is unknown for subjects assigned control who are not canvassed: these subjects form a mixture of compliers and never-takers. For each individual in the study, $K = 4$ pretreatment covariates are available: age, party affiliation (Democrat or Republican), abstention from the 1996 election (yes or no), and voting in the 1996 election (yes or no), denoted as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$. All four covariates are used to model strata membership and potential outcomes distributions, with models similar to (4.3) and (4.4). The final vector of model parameters, $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\beta})$, has 13 components. See Appendix C.2 for more details on models and computations.

We apply the HMC-within-Gibbs algorithm as well as a standard MHMC-within-Gibbs algorithm to obtain draws form the posterior distribution of model parameters $\boldsymbol{\theta}$ and estimands (4.6), and compare their performance. The specific MHMC algorithm considered is similar to the HMC sampler, with the exception that parameter draws are obtained by a random walk Metropolis step instead of an HMC step, within the overall Gibbs procedure. Also, the covariance matrix of the Normal proposal equals to the matrix inverse of the Hessian of the negative logarithm of the current conditional posterior.

We ran 10 independent chains of length 2000 with random initializations and discarded a burn-in of 1000 draws. Table 4.6 summarizes diagnostic statistics, calculated for the obtained draws, i.e., GR statistics (Gelman and Rubin 1992) and effective sample sizes (ESS). It is evident from the table that the HMC algorithm is superior to MHMC for this problem in all parameters, i.e., it produces GR statistics that are closer to one and results in larger ESS. In addition, Figures 4.2 and 4.3 show autocorrelation plots of draws generated using HMC and MHMC, respectively, for

Table 4.6: Summary of diagnostics for HMC-within-Gibbs and MHMC-within-Gibbs algorithms, applied to Example 1 described in Section 4.4.2. Here, π_ψ and π_β denote log-posteriors for each set of parameters. Note that HMC outperformed MHMC on all parameters under each setting.

	HMC			
	Without Exclusion		With Exclusion	
Parameter	GR	ESS	GR	ESS
π_ψ	1.01	1724	1.01	1798
π_β	1.07	117	1.06	752
ITT_C	1.21	80	1.02	644
ITT_N	1.20	96	1.01	1865
	MHMC			
	Without Exclusion		With Exclusion	
Parameters	GR	ESS	GR	ESS
π_ψ	1.13	211	1.12	184
π_β	2.53	72	1.42	123
ITT_C	3.06	101	1.32	205
ITT_N	2.68	169	1.09	1223

the case with exclusion restrictions. The gain in efficiency is apparent from the fact that autocorrelations between consecutive and near-consecutive draws are generally smaller for the HMC.

4.4.3 Example 2: Influenza Vaccination and Flu

For our second example we use data from a study of a causal effect of influenza vaccination on flu-related hospitalization visits, described in Hirano et al. (2000). The experiment consisted of sending letters to randomly chosen group of doctors, encouraging them to inoculate their patients. However, a patient could choose to ignore doctor's encouragement to get a vaccination.

Let z_i indicate whether the patient i was encouraged ($z_i = 1$) or not ($z_i = 0$). Let $d_i(z)$ indicate whether subject i received a flu vaccine under treatment z. There

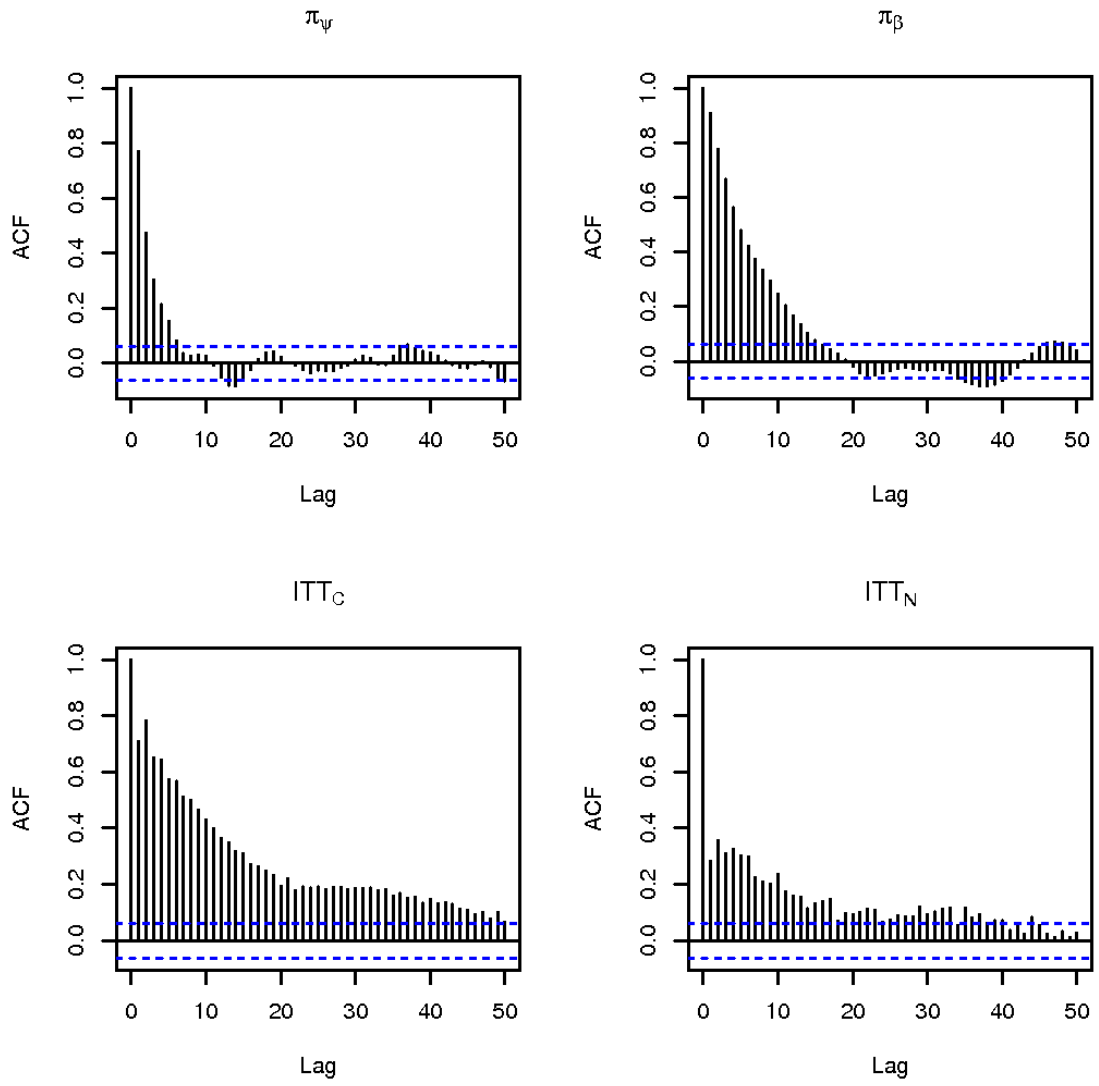


Figure 4.2: Autocorrelation plots of draws produced by HMC algorithm in Example 1, described in Section 4.4.2, for the case with exclusion restrictions. They show relatively low autocorrelation between consecutive and near-consecutive draws.

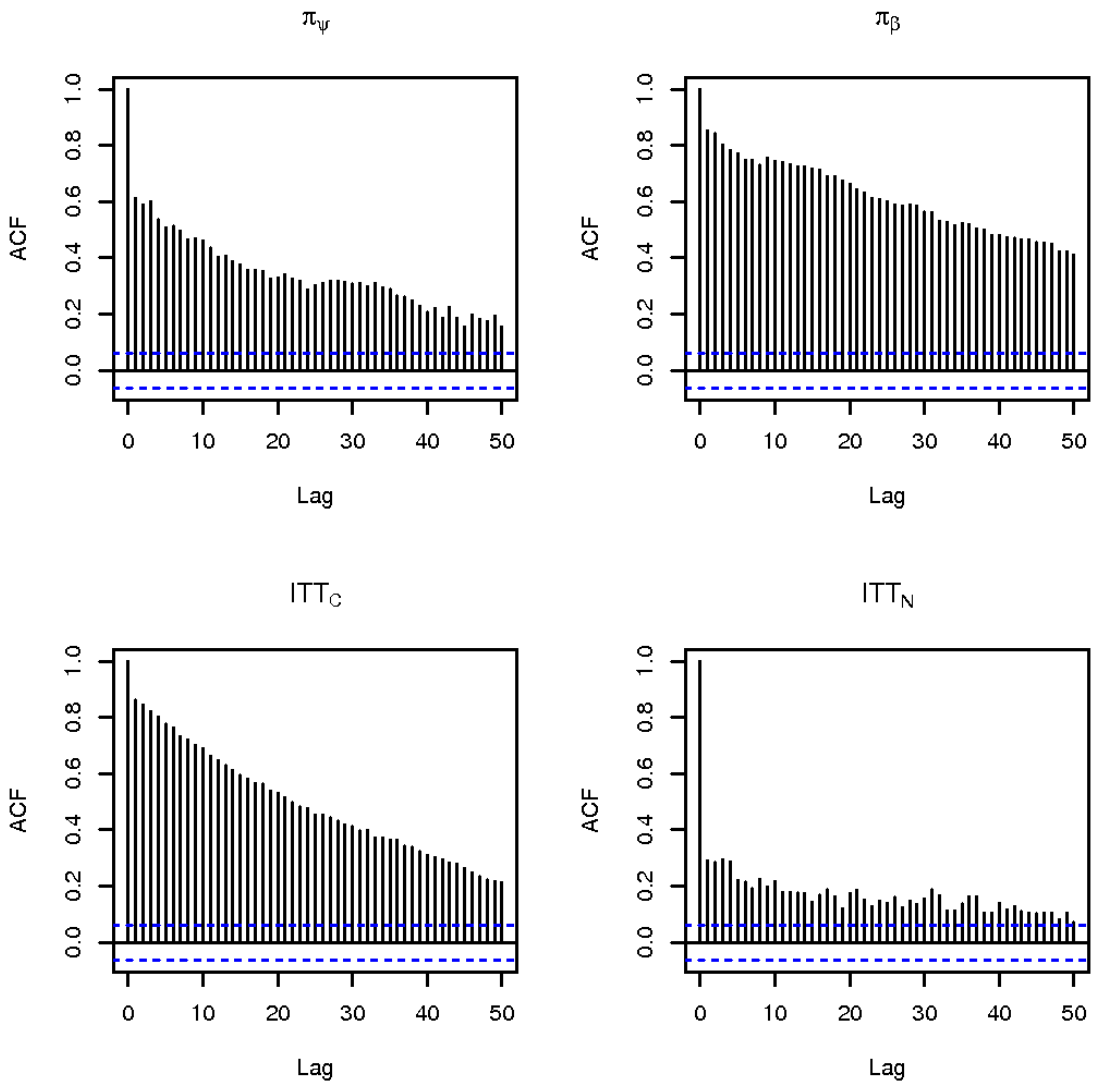


Figure 4.3: Autocorrelation plots of draws produced by MHMC algorithm in Example 1, described in Section 4.4.2, for the case with exclusion restrictions. As compared to the plots in Figure 4.2, correlations between consecutive and near-consecutive draws are much higher.

are three principal strata in this experiment: compliers ($d_i(z) = z$), never-takers ($d_i(0) = d_i(1) = 0$), and always-takers ($d_i(0) = d_i(1) = 1$), represented by $s_i = c, n$, and a, respectively. Potential outcome $y_i(z)$ is defined as

$$y_i(z) = \begin{cases} 1 & \text{if subject } i \text{ had a flu-related hospitalization under treatment } z, \\ 0 & \text{otherwise.} \end{cases}$$

For a total $N = 2,891$ experimental subjects, the notation and model setup are similar to the example in Section 4.4.2, with the difference that only $K = 2$ pre-treatment covariates are recorded for each subject: age, and an indicator for chronic obstructive pulmonary disease. The goal is to estimate model parameters $\boldsymbol{\theta}$, 13 in total, and three estimands of interest: finite population causal effects defined as in (4.6) for each of three strata. Further details are outlined in Appendix C.3.

Once again, we compare the performance of our HMC-within-Gibbs algorithm to that of a standard MHMC algorithm, and summarize the results in Table 4.7. HMC outperforms MHMC in GR statistics for all parameters. However, comparison of ESS indicates that, although in majority of settings and for most of the parameters HMC has higher ESS, the results are not as consistent for the estimates of ITT_A and ITT_N. Examples of autocorrelation plots are included in Appendix C.3.

4.5 Results and Discussion

We apply HMC-within-Gibbs algorithm to the data from a clinical trial under consideration to estimate the estimands of interest described in Section 4.3.2. The inference is complicated by a small sample size, large number of parameters, and

Table 4.7: Summary of diagnostics for HMC-within-Gibbs and MHMC-within-Gibbs algorithms, applied to Example 2 described in Section 4.4.3. Symbol "00" denotes no exclusion restrictions, "01" denotes an exclusion restriction on always-takers only, "10" denotes an exclusion restriction on never-takers only, and "11" denotes exclusion restrictions on both never-takers and always-takers. Also, π_ψ and π_β denote log-posteriors for each set of parameters. The HMC algorithm is superior for the vast majority of estimates.

	HMC							
Exclusion type	00		01		10		11	
Parameters	GR	ESS	GR	ESS	GR	ESS	GR	ESS
π_ψ	1.02	735	1.03	764	1.05	736	1.03	756
π_β	1.48	67	1.60	145	1.07	456	1.01	1173
ITT_C	1.28	75	1.45	125	1.04	273	1.05	547
ITT_N	1.35	64	1.43	126	1.03	697	1.01	1279
ITT_A	1.08	157	1.01	3239	1.09	214	1.02	3623
	MHMC							
Exclusion type	00		01		10		11	
Parameters	GR	ESS	GR	ESS	GR	ESS	GR	ESS
π_ψ	1.32	117	1.17	172	1.39	117	1.47	131
π_β	3.01	46	5.14	136	1.60	110	1.19	248
ITT_C	3.5	55	4.49	80	1.58	67	1.20	95
ITT_N	3.15	161	4.39	291	1.05	455	1.07	454
ITT_A	2.65	337	1.05	2595	2.16	227	1.04	2173

missingness in the outcomes. Therefore, we do not attempt to use MHMC algorithm on these data. Appendix C.1 outlines the computational details, including Gibbs steps and HMC specifics for posterior computations.

The resulting estimates are summarized in Table 4.8. Overall, the conclusion from the analysis is the same as the one obtained initially, i.e., the treatment and control groups do not differ in the rates of adverse events. However, our current analysis is more accurate because it estimates treatment effects for subsets of patients for which both potential outcomes are well-defined.

Table 4.9 shows estimated average percentages of units per each strata. Although

Table 4.8: Posterior median and 95% posterior intervals for estimands of interest listed in Section 4.3.2, estimated for the clinical trial under consideration using HMC algorithm. As expected, there is no indication of increased rate of adverse events in the treatment group, as compared to the rates in the control group. In addition, PS method allowed us to estimate the rates of death under the treatment and under the control at each of two time points.

Parameter	Median	95% Posterior Interval
δ_1	-0.03	$(-0.40, 0.21)$
δ_2	-0.01	$(-0.24, 0.18)$
ξ_1^0	0.17	$(0.03, 0.21)$
ξ_1^1	0.01	$(0, 0.03)$
ξ_2^0	0.04	$(0.01, 0.18)$
ξ_2^1	0.19	$(0.17, 0.21)$

Table 4.9: Average percentage of units estimated for each strata defined in Table 4.4.

aa	*pa*	*na*	*pp*	*np*	*nn*
0.790	0.005	0.077	0.004	0.115	0.009

in the initial analysis of these data, described in Chapter 4.9, it was essentially assumed that all subjects belonged to strata *aa*, the results obtained here indicate that an estimated 20% of subjects would not survive until the end of the study under one (or both) of the treatment(s).

To summarize, in this chapter we reviewed PS framework and the attendant assumptions. Then we introduced the HMC-within-Gibbs algorithm that is especially suitable for sampling from the posterior distribution of the parameters in PS, which is usually complicated by weak identifiability of the model. We demonstrated the superiority of the HMC algorithm over a traditional MHMC-within-Gibbs method on two real-data examples, and, finally, we applied the described method to the data from a medical device clinical trial. Although the actual conclusion has not changed, the analysis showed that there is evidence that the data support the PS model.

Chapter 5

Conclusion

Nearly a century-long effort of developing and studying methods of handling missing data had a major breakthrough in 1970th, when in a series of publications by D. B. Rubin and other statisticians it was proposed to treat missingness as a random process. This idea helped to formalize the problem of dealing with missing data by reformulating it as a problem of modeling missingness mechanism. A plethora of new methods has been proposed since then, transforming the ways that missing data are treated.

With the formalization of the problem came the realization that almost all methods of missing data handling rely on unassessable assumptions about the nature of the missingness mechanism. Many recently issued guidelines for handling missing data emphasize the need for standardization of requirements for reporting missing data and methods of conducting sensitivity analyses in empirical studies (Burzykowski et al. 2010; CHMP 2010; NRC-Panel 2010). However, recommendations for specific features and characteristics of missing data that have to be reported as well as for

methods of performing systematic sensitivity analyses are still scarce, and there is little consensus in the field. In Chapter 1 we gave an overview of modern classification of missing data mechanisms and methods for parameter estimation for incomplete data. We also provided some recommendations on reporting missing data, including informative summaries and graphical representations that should be part of every experimental or observational study report.

In Chapters 2 and 3 we developed a general method of sensitivity analysis of study's conclusions to assumptions about missing data. The method uses graphical displays to demonstrate sensitivity of the estimate of the treatment effect to alternative missing data specifications and to identify tipping points of the study. In Chapter 2 we described a basic version of enhanced TP displays that help visualize the results of sensitivity analyses for studies with binary outcomes. In Chapter 3 we generalized this idea and proposed a systematic way of performing sensitivity analyses based on a pattern-mixture decomposition of a joint model for outcomes and missingness indicators. In addition, we presented a series of sensitivity parameters that can be used to explore alternative models for the missingness mechanism and to assess the strength of the study's conclusions.

Finally, in Chapter 4, we described another method of performing sensitivity analyses using PS framework. We also proposed an improved method of computation using HMC algorithm, which accelerates posterior calculations under the PS. All together, the proposed approaches form a novel collection of useful tools for the analysis of data sets plagued with missing values.

Appendix A

Missing Data Handling

A.1 Violation of Distinctness Under MAR

Let $\boldsymbol{Y} = (y_1, \ldots, y_N)'$, $\boldsymbol{X}_{obs} = \boldsymbol{X}$ and $y_i = \theta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and σ is known. Let's consider two types of censoring of outcomes $\boldsymbol{Y}$. For the first one, all units with $y_i < c$, where c is a know constant, will be considered missing, then $\boldsymbol{Y}_{obs} = \{y_i : y_i < c\} = (y_{m_1}, y_{m_2}, \ldots, y_{m_r})'$, $r \leq N$. For the second one, $\tilde{\boldsymbol{Y}}_{obs} = \{y_i : \theta x_i < c\} = (y_{n_1}, y_{n_2}, \ldots, y_{n_{\tilde{r}}})'$, $\tilde{r} \leq N$. Also, let $\boldsymbol{D}$ and $\tilde{\boldsymbol{D}}$ be the corresponding vectors of missingness indicators.

On the surface, these two mechanisms are quite similar, however a closer look reveals fundamental differences. The following is the joint density for the observed

data in the first case:

$$\begin{aligned} f(\boldsymbol{Y}_{obs}, \boldsymbol{D} \mid \boldsymbol{X}, \theta) &= \prod_{i=1}^{r} f(y_{m_i}, d_{m_i} \mid x_{m_i}, \theta) \prod_{i=r+1}^{N} f(d_{m_i} \mid x_{m_i}, \theta) \\ &= \prod_{i=1}^{r} f(y_{m_i} \mid x_{m_i}, \theta) P(y_{m_i} \geq c \mid x_{m_i}, y_{m_i}, \theta) \prod_{i=r+1}^{N} P(y_{m_i} < c \mid x_{m_i}, \theta) \\ &= \prod_{i=1}^{r} \phi\left(\frac{y_{m_i} - \theta x_{m_i}}{\sigma}\right) \prod_{i=r+1}^{N} \Phi\left(\frac{y_{m_i} - \theta x_{m_i}}{\sigma} < \frac{c - \theta x_{m_i}}{\sigma}\right), \end{aligned}$$

where indexes m_{r+1} through m_N correspond to censored units. Last equality holds because, for $m_i \leq r$ (respondents), $P(y_{m_i} \geq c | x_{m_i}, y_{m_i}, \theta) = 1$. Here $\phi(\cdot)$ is a standard Normal probability distribution function, and $\Phi(\cdot)$ is a corresponding cumulative distribution function. Clearly, the second product, which models the missing data mechanism, can not be dropped from the likelihood of θ. Therefore, MLE estimate will not correspond to an estimate obtained by regressing observed outcomes on $\boldsymbol{X}$ (i.e., CCA). Another way to look at it is that censoring based on values of $\boldsymbol{Y}$ leads to a violation of a fundamental assumption of normality in the linear regression, because $(y_{m_i} - \theta x_{m_i}) | \theta, x_{m_i}$ is now distributed as a *truncated* normal $N_{c+}(0, \sigma^2)$.

On the other hand, under the second scenario,

$$\begin{aligned} f(\tilde{\boldsymbol{Y}}_{obs}, \tilde{\boldsymbol{D}} \mid \boldsymbol{X}, \theta) &= \prod_{i=1}^{\tilde{r}} f(y_{n_i} \mid d_{n_i}, x_{n_i}, \theta) \prod_{i=\tilde{r}+1}^{N} f(d_{n_i} \mid x_{n_i}, \theta) \\ &= \prod_{i=1}^{\tilde{r}} f(y_{n_i} \mid x_{n_i}, \theta) P(\theta x_{n_i} \geq c \mid x_{n_i}, y_{n_i}, \theta) \prod_{i=\tilde{r}+1}^{N} P(\theta x_{n_i} < c \mid x_{n_i}, \theta) \\ &= \prod_{i=1}^{\tilde{r}} \phi\left(\frac{y_{n_i} - \theta x_{n_i}}{\sigma}\right) I(\theta x_{n_i} \geq c) \prod_{i=\tilde{r}+1}^{N} I(\theta x_{n_i} < c). \end{aligned}$$

Here, the conditional distribution of $\tilde{\boldsymbol{Y}}_{obs}$ given $\boldsymbol{X}$ is Normal and the standard CCA

analysis would produce unbiased estimates of θ.

However, there is a way to gain more efficiency in the second scenario. Notice that $\theta = \phi$, i.e., it is a special case of nonignorable missing data with MAR mechanism. Since parameters are not distinct, we can get more precise estimates of θ by maximizing the likelihood $L(\theta \mid \tilde{\boldsymbol{Y}}_{obs}) = \prod_{i=1}^{\tilde{r}} \phi\left(\frac{y_{n_i} - \theta x_{n_i}}{\sigma}\right)$ and use the following N constrains to improve the precision of the estimate for θ,

$$\begin{cases} \theta x_{n_i} \geq c \text{ for } i \geq \tilde{r}, \\ \theta x_{n_i} < c \text{ for } i < \tilde{r}. \end{cases}$$

As an immediate corollary, we can notice that the *smaller* the error σ is the closer these two quantities are $P(y_i > c) \approx P(\theta x_i > c)$. Therefore, the better the predictive model is for $\boldsymbol{Y} \mid \boldsymbol{X}, \theta$, the closer the (possibly) MNAR mechanism may be approximated by MAR.

Appendix B

ETP Displays

B.1 Minimal Sufficiency for EF and NEF

Many regularly used distributions belong to a class of *exponential families* (EF).

Definition B.1.1. *The distribution of a random variable Y is a member of an EF of order one (EF_1) if its density has the following form*

$$f(y) = \exp\{s(y)\eta(\theta) - b(\theta) + c(y)\},$$

where θ is a scalar parameter.

Here, $\eta(\cdot)$ is a function of θ called the *natural parameter*, $s(y)$ is called *natural observation*, and $\psi(\eta) \equiv b(\theta)$ is a cumulant function. If $s(y)$ is linear, the family becomes a *natural EF_1* (NEF_1). For an i.i.d. sample $y_1, y_2, \ldots, y_N$ from an NEF_1

with $s(y) = y$, the likelihood for θ is proportional to

$$\exp(\eta(\theta)\sum y_i - Nb(\theta)).$$

Therefore, $\sum y_i$ (or $\bar{y}$) is a minimal sufficient statistic (MSS) for θ. As discussed in Section 3.2, if Y is the outcome of interest and its distribution is of an NEF_1 type, then the use of ETP displays for the purpose of sensitivity analyses of the treatment effect to missing data becomes straightforward, because the one-dimensional MSS provides a natural data summary that can be represented by horizontal and vertical axes.

Some commonly used distributions are members of a particular subclass of NEF_1 with quadratic variance function (NEF-QVF, Morris 1982, 1983; Morris and Lock 2009). If Y is distributed as NEF-QVF with $\mu \equiv E(Y \mid \theta) = \psi'(\eta)$, then $Var(Y \mid \theta) = \psi''(\eta) = \upsilon_2\mu^2 + \upsilon_1\mu + \sigma_0$, where υ_1, υ_2 and σ_0 are known. This class includes the following six distributions:

- Normal distribution with known variance σ_0, $N(\mu, \sigma_0)$, $\eta = \mu$;
- Poisson distribution $\text{Pois}(\mu)$, $\eta = \log(\mu)$;
- Exponential distribution with scale parameter μ, $\mu\text{Expo}(1)$, and a scaled Gamma distribution with known shape parameter α, $\mu\text{Gam}(\alpha)$, with $\eta = 1 - 1/\mu$;
- Bernoulli distribution $\text{Bern}(p)$, with $p = \mu$, and $\text{Binom}(N, p)$ with known N and $\eta = \log(p/(1-p))$;

- Geometric distribution Geom(p), with $\mu = p/(1-p)$ and $\eta = \log(2p)$, and Negative binomial NegBinom(r, p) with known convolution parameter r;
- Less commonly used, but still quite handy, Convolved Hyperbolic Secant distribution, CHS(μ); it is skewed with support on the real line and $\eta = \tan^{-1}(\mu)$.

Many applied modeling problems can be addressed by employing EFs. However, as seen from the Definition B.1.1, every EF_1 is a non-linear transformation of a corresponding NEF_1, e.g., Lognormal, Weibull, Pareto, Chi, Power Function, Inverted Gamma etc. If the distribution of Y is EF_1 and $s(y)$ is a monotone function, then the problem can be reduced back to the NEF_1 case by working with a transformed variable $z = s(y)$, so that the MSS for the problem becomes $\sum z_i$. This expands the pool of distributions for which the ETP displays can be used straightaway.

Finally, the preceding discussion can be generalized to models with several parameters by defining an EF of order p (EF_p).

Definition B.1.2. *The distribution of a random variable Y is a member of EF_p if its density has the following form*

$$f(y) = \exp\{\boldsymbol{s}(y)^T\boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(y)\},$$

where $\boldsymbol{s}(y) = (s_1(y), s_2(y), \ldots, s_p(y))'$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_p)'$ *and* $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$.

If p, which is also the dimension of the sufficient statistic, does not match q, the dimension of the parameter-vector, then the family is called *curved EFs*. Models with $q > p$ are generally not useful because, in this case, $\boldsymbol{\theta}$ can not be identified from the

data. If $p = q$ and $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a 1-to-1 mapping then the MSS for $\boldsymbol{\theta}$ is

$$(\sum_{i=1}^{N} s_1(y_i), \sum_{i=1}^{N} s_2(y_i), \ldots, \sum_{i=1}^{N} s_p(y_i)).$$

This provides a natural way to generalize ETP displays to problems with outcomes modeled as an $\mathrm{EF_p}$. It can be done by fixing some components of the MSS while plotting the others, as illustrated in Section 3.2.1 for a Normal model with unknown mean and variance.

However, if the parameter of interest is a component of a multiple regression coefficient $\boldsymbol{\beta}$ from $f(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta})$, the choice of the convenient summary is less apparent. Given a canonical link function, the MSS for $\boldsymbol{\beta}$ in GLM is a vector $\boldsymbol{X}^T\boldsymbol{Y}$ (McCullagh and Nelder 1989), and further research is needed to find the best and most intuitive way to reduce this MSS to a one-dimensional summary.

B.2 Approximate Degrees of Freedom

The proof of Theorem 3.2.2 is based on matching first two moments of the distribution of the squared denominator in (3.4) to a scaled chi-square distribution. Let $s_d^2 = \frac{s_1^2}{N^T} + \frac{s_0^2}{N^C}$, then

$$s_d^2 \mid \sigma_0^2, \sigma_1^2 \sim \frac{\sigma_1^2 \chi^2_{N^T_{obs}}}{N^T N^T_{obs}} + \frac{\sigma_0^2 \chi^2_{N^C_{obs}}}{N^C N^C_{obs}},$$

where $\chi^2_{N^T_{obs}}$ and $\chi^2_{N^C_{obs}}$ are two independent chi-square distributions. The mean and variance of s_d^2 are the following

$$E(s_d^2 \mid \sigma_0^2, \sigma_1^2) = \sigma_1^2/N^T + \sigma_0^2/N^C, \; Var(s_d^2 \mid \sigma_0^2, \sigma_1^2) = 2\left((\sigma_1^2/N^T)^2/N^T_{obs} + (\sigma_0^2/N^C)^2/N^C_{obs}\right).$$

Let's consider a new chi-square distribution, with scale parameter $\tilde{s}^2$ and degrees of freedom $\tilde{f}$, that has the same first two moments as s_d^2. Then,

$$\begin{aligned}\tilde{s}^2\tilde{f} &= \sigma_1^2/N^T + \sigma_0^2/N^C,\\ \tilde{s}^4\tilde{f} &= (\sigma_1^2/N^T)^2/N_{obs}^T + (\sigma_0^2/N^C)^2/N_{obs}^C.\end{aligned}$$

It follows that

$$\tilde{f} = \frac{(\sigma_1^2/N^T + \sigma_0^2/N^C)^2}{(\sigma_1^2/N^T)^2/N_{obs}^T + (\sigma_0^2/N^C)^2/N_{obs}^C}.$$

Finally, $\hat{\tilde{f}}$ in (3.5) is obtained by substituting σ_i^2 with s_i^2, $i = 0, 1$. The discussion of the validity of the test under the derived approximation presented in Welch (1938) applies to the current modification.

Appendix C

HMC Algorithm for PS Framework

C.1 Bayesian Updating for PS Framework with HMC Steps

The sampler described below calculates the posterior of estimands of interest by iterating between imputing missing principal strata conditional on the current set of parameter draws, then drawing from the posterior distribution of the parameters conditional on the imputed strata using HMC, and finally using the results from the previous two steps to impute missing outcomes and calculate the estimands of interest.

The model for the clinical trial under consideration was set up in Section 4.3.3. A series of assumptions and simplifications resulted in the following vector of parame-

ters,

$$\begin{aligned}\boldsymbol{\theta} =&(\beta_{aa,110}, \beta_{pa,110}, \beta_{pp,110}, \beta_{aa,111}, \beta_{pa,111}, \beta_{na,111},\\ &\beta_{pp,111}, \beta_{np,111}, \beta_{aa,120}, \beta_{aa,121}, \beta_{pa,121}, \beta_{na,121},\\ &\beta_{aa,210}, \beta_{pa,210}, \beta_{pp,210}, \beta_{aa,211}, \beta_{pa,211}, \beta_{na,211},\\ &\beta_{pp,211}, \beta_{np,211}, \beta_{aa,220}, \beta_{aa,221}, \beta_{pa,221}, \beta_{na,221},\\ &\beta_{aa,310}, \beta_{pa,310}, \beta_{pp,310}, \beta_{aa,311}, \beta_{pa,311}, \beta_{na,311},\\ &\beta_{pp,311}, \beta_{np,311}, \beta_{aa,320}, \beta_{aa,321}, \beta_{pa,321}, \beta_{na,321},\\ &\beta_1, \beta_2, \psi_{aa}, \psi_{pa}, \psi_{pp}, \psi_{na}, \psi_{np}, \psi_1, \psi_2, \psi_3, \psi_4).\end{aligned}$$

Here, to keep the notation uncluttered, we renamed $\psi_h \equiv \psi_{h,0}$ and $\beta_{h,rtz} \equiv \beta_{h,rtz,0}$. Let

$$\begin{aligned}\Psi_s(\boldsymbol{x}_i) &= \frac{\exp(\psi_{s,0} + \psi_1 x_{i1} + \psi_2 x_{i2} + \psi_3 x_{i3} + \psi_4 x_{i4})}{\sum_{h\in\Omega} \exp\left(\psi_{h,0} + \psi_1 x_{i1} + \psi_2 x_{i2} + \psi_3 x_{i3} + \psi_4 x_{i4}\right)},\\ \Gamma(\boldsymbol{x}_i, \beta_{s,rtz}) &= \frac{\exp\left(\beta_{s,rtz} + \beta_1 x_{i5} + \beta_2 x_{i6}\right)}{1 + \exp\left(\beta_{s,rtz} + \beta_1 x_{i5} + \beta_2 x_{i6}\right)}.\end{aligned}$$

Also, let

$$\begin{aligned}\Phi_s(i,t) =&\Gamma(\boldsymbol{x}_i, \beta_{s,1tz_i})^{y_{1i,t}(z_i)} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{s,1tz_i})\right)^{5-y_{1i,t}(z_i)} \times\\ &\Gamma(\boldsymbol{x}_i, \beta_{s,2tz_i})^{y_{2i,t}(z_i)} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{s,2tz_i})\right)^{1-y_{2i,t}(z_i)} \times\\ &\Gamma(\boldsymbol{x}_i, \beta_{s,3tz_i})^{y_{3i,t}(z_i)} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{s,3tz_i})\right)^{1-y_{3i,t}(z_i)}.\end{aligned}$$

Then, the likelihood in (4.2) is proportional to the following:

$$\Phi = \prod_{i:s_i=aa} \Psi_{aa}(\boldsymbol{x}_i)\Phi_{aa}(i,1)\Phi_{aa}(i,2) \cdot \prod_{i:s_i=pa} \Psi_{pa}(\boldsymbol{x}_i)\Phi_{pa}(i,1)\left[\Phi_{pa}(i,2)\right]^{z_i} \cdot \tag{C.1}$$
$$\prod_{i:s_i=pp} \Psi_{pp}(\boldsymbol{x}_i)\Phi_{pp}(i,1) \cdot \prod_{i:s_i=na} \Psi_{na}(\boldsymbol{x}_i)\left[\Phi_{na}(i,1)\Phi_{na}(i,2)\right]^{z_i} \cdot$$
$$\prod_{i:s_i=np} \Psi_{np}(\boldsymbol{x}_i)\left[\Phi_{np}(i,1)\right]^{z_i} \cdot \prod_{i:s_i=nn} \Psi_{nn}(\boldsymbol{x}_i).$$

We assume that all parameters are a priori independent and follow Normal distributions with mean 0 and standard deviation 2.5. Also, all continuous covariates are standardized to have mean 0 and standard deviation 2.5.

Sampling from the posterior, which is proportional to C.1, is fairly straightforward. The estimation algorithm consists of the following steps:

0. Initialize latent strata $\boldsymbol{S}$;

1. Update parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\beta})$ given latent strata using HMC. Conditional on $\boldsymbol{S}$, the vectors of parameters $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ are independent *a posteriori*, and the gradient of the logarithm of each posterior can be obtained in closed-form. Suppose the current draw is $\boldsymbol{\theta}^{(t)}$. The HMC algorithm to sample each $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ separately consists of the following steps:

   - Sample new momentum vector $\boldsymbol{p}(0)$ from a multivariate Gaussian distribution.
   - Perform L steps of the *Leapfrog* algorithm Hockney (1970), starting at $\boldsymbol{\theta}^{(t)}(0) \equiv \boldsymbol{\theta}^{(t)}$ and $\boldsymbol{p}(0)$, to obtain a new proposal $(\boldsymbol{\theta}^{(t)}(L), \boldsymbol{p}(L))$ for the augmented parameters space. The algorithm uses the following approximation

to update parameters (position) $\boldsymbol{\theta}$ and momentum $\boldsymbol{p}$:

$$\begin{aligned}
\boldsymbol{p}(t+\epsilon/2) &= \boldsymbol{p}(t) - \frac{\epsilon}{2}\frac{\partial H(\boldsymbol{\theta},\boldsymbol{p})}{\partial \boldsymbol{\theta}}\bigg|_{t}, \\
\boldsymbol{\theta}(t+\epsilon) &= \boldsymbol{\theta}(t) + \epsilon \boldsymbol{p}(t+\epsilon/2)\Lambda^{-1}, \\
\boldsymbol{p}(t+\epsilon) &= \boldsymbol{p}(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial H(\boldsymbol{\theta},\boldsymbol{p})}{\partial \boldsymbol{\theta}}\bigg|_{t+\epsilon}.
\end{aligned}$$

- Accept the new proposal, i.e. let $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}(L)$, with the probability

$$\min\{1, \frac{\exp\left(H(\boldsymbol{\theta}^{(t)}(0), \boldsymbol{p}(0))\right)}{\exp\left(H(\boldsymbol{\theta}^{(t)}(L), \boldsymbol{p}(L))\right)}\},$$

  otherwise, let $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$. This rule is similar to the one used in MHMC, except that here the acceptance probability depends on the ratio of "energies" at the current state and at the end state of the leap-frog path.

2. Impute latent strata given updated parameters. The distribution of $\boldsymbol{S}$ conditional on observed data and model parameters $\boldsymbol{\theta}$ is easy to calculate by Bayes' theorem. For that, we use Table 4.5 to identify what strata are possible for each subject, given their treatment group and observe outcomes. If there is more than one possible stratum, then we use multinomial distribution to select it probabilistically. For example, if $t_i = 1$ and $d_{i,1} = d_{i,2} = 0$, then the subject can belong to stratum aa, pa, or na. In order choose one, we sample from

Multinom $\left(1, \left(\frac{\rho_{aa}}{\rho}, \frac{\rho_{pa}}{\rho}, \frac{\rho_{na}}{\rho}\right)\right)$, where

$$\begin{aligned}
\rho_{aa} &= \Psi_{aa}(\boldsymbol{x}_i)\Phi_{aa}(i,1)\Phi_{aa}(i,2), \\
\rho_{pa} &= \Psi_{pa}(\boldsymbol{x}_i)\Phi_{pa}(i,1)\Phi_{pa}(i,2), \\
\rho_{na} &= \Psi_{na}(\boldsymbol{x}_i)\Phi_{na}(i,1)\Phi_{na}(i,2), \\
\rho &= \rho_{aa} + \rho_{pa} + \rho_{na}.
\end{aligned}$$

3. Given the values of parameters and sampled strata, impute missing outcomes using models 4.4 and 4.5. Steps 1-3 are iterated until convergence.

4. Estimate the estimands of interest: $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, ξ_1^z, and ξ_2^z, $z \in \{0, 1\}$.

We set the HMC leapfrog step size to $\epsilon_\psi = 0.05$ and take $L_\psi = 10$ leapfrog steps for parameters $\boldsymbol{\psi}$, and also set $\epsilon_\beta = 0.05$ and $L_\beta = 10$ for parameters $\boldsymbol{\beta}$. We produce three chains of length 100,000 and discard first 30% burn-in draws.

C.2 Data and Models for Example 1

Table C.1 summarizes the observed data used for Example 1 in Section 4.4.2.

The following assumptions on the potential outcomes were employed,

$$y_i(0) \perp\!\!\!\perp y_i(1) \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \tag{C.2}$$

$$z_i \mid y_i(0), y_i(1), d_i(0), d_i(1), d_i, \boldsymbol{x}_i, \boldsymbol{\theta} \;\sim\; z_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \tag{C.3}$$

Table C.1: Observed outcomes in the Gerber and Green (2000) field trial. The data is limited to single-occupancy homes, with individuals assigned to receive the personal canvassing treatment (and no other treatments) or those assigned no treatment at all.

Assignment z_i^{obs}	Canvassed d_i^{obs}	Voted y_i^{obs}	# Subjects ($N = 6617$)	Strata
0	0	0	3168	c or n
0	0	1	2101	c or n
1	0	0	595	c
1	0	1	381	n
1	1	0	163	c
1	1	1	209	c

$$\begin{aligned}
\Psi(\boldsymbol{x}_i) &= \Pr\left(s_i = c \mid \boldsymbol{x}_i; \boldsymbol{\theta}\right) = 1 - \Pr(s_i = n \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \\
&= \frac{\exp(\psi_0 + \psi_1 x_{i1} + \psi_2 x_{i2} + \psi_3 x_{i3} + \psi_4 x_{i4})}{1 + \exp(\psi_0 + \psi_1 x_{i1} + \psi_2 x_{i2} + \psi_3 x_{i3} + \psi_4 x_{i4})}, \\
\Gamma(\boldsymbol{x}_i, \beta_{sz0}) &= \Pr\left(y_i(z_i) = 1 \mid s_i = s, z_i = z, \boldsymbol{x}_i; \boldsymbol{\theta}\right) \\
&= \frac{\exp(\beta_{sz0} + \beta_{0\cdot1} x_{i1} + \beta_{0\cdot2} x_{i2} + \beta_{0\cdot3} x_{i3} + \beta_{0\cdot4} x_{i4})}{1 + \exp(\beta_{sz0} + \beta_{0\cdot1} x_{i1} + \beta_{0\cdot2} x_{i2} + \beta_{0\cdot3} x_{i3} + \beta_{0\cdot4} x_{i4})}.
\end{aligned}$$

The assumption on the assignment mechanism (C.3) is justified by the study design. There are a total of 13 parameters in this model,

$$\boldsymbol{\theta} = (\psi_0, \psi_1, \psi_2, \psi_3, \psi_4, \beta_{c00}, \beta_{c10}, \beta_{n00}, \beta_{n10}, \beta_{0\cdot1}, \beta_{0\cdot2}, \beta_{0\cdot3}, \beta_{0\cdot4}).$$

We assume that intercepts and slopes follow Cauchy distributions with scale 2.5 independently a priori. Continuous covariates are standardized to have mean 0 and standard deviation 2.5. The crucial piece of missing data in this study are compliances for subjects assigned control who receive control.

The complete-data likelihood has the following form:

$$\begin{aligned}
&\prod_{i:s_i=c}\Big[\prod_{i\in(0,0)}\left\{\Psi(\boldsymbol{x}_i)\times\Gamma(\boldsymbol{x}_i,\beta_{c00})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{c00})\right)^{1-y_i}\right\}\\
&\qquad\prod_{i\in(1,0)}\left\{\Psi(\boldsymbol{x}_i)\times\Gamma(\boldsymbol{x}_i,\beta_{c10})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{c10})\right)^{1-y_i}\right\}\Big]\times\\
&\prod_{i:s_i=n}\Big[\prod_{i\in(0,0)}\left\{(1-\Psi(\boldsymbol{x}_i))\times\Gamma(\boldsymbol{x}_i,\beta_{n00})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{n00})\right)^{1-y_i}\right\}\times\\
&\qquad\prod_{i\in(1,0)}\left\{(1-\Psi(\boldsymbol{x}_i))\times\Gamma(\boldsymbol{x}_i,\beta_{n10})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{n10})\right)^{1-y_i}\right\}\Big],
\end{aligned}\tag{C.4}$$

where $y_i = y_i(1)d_i(z_i) + y_i(0)\{1-d_i(z_i)\}$, and $i \in (z,d)$ if subject i is assigned treatment z and receives treatment d, where $z, d \in \{0,1\}$.

The distribution of $\boldsymbol{S}$ conditional on observed data and model parameters $\boldsymbol{\theta}$ is easily derived. For example, if a subject assigned control receives control, $d_i(0) = 0$, then the conditional probability that the subject is a complier is

$$\Pr\left(s_i = c \mid d_i(0) = 0, \boldsymbol{x}_i; \boldsymbol{\theta}\right) =$$

$$\frac{\Psi(\boldsymbol{x}_i)\Gamma(\boldsymbol{x}_i,\beta_{c00})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{c00})^{1-y_i}\right)}{\Psi(\boldsymbol{x}_i)\Gamma(\boldsymbol{x}_i,\beta_{c00})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{c00})^{1-y_i}\right)+(1-\Psi(\boldsymbol{x}_i))\Gamma(\boldsymbol{x}_i,\beta_{n00})^{y_i}\left(1-\Gamma(\boldsymbol{x}_i,\beta_{n00})^{1-y_i}\right)}$$

We set the HMC leapfrog step size to $\epsilon_\psi = 0.04$ and take $L_\psi = 100$ leapfrog steps for parameters $\boldsymbol{\psi}$, and also set $\epsilon_\beta = 0.03$ and $L_\beta = 50$ for parameters $\boldsymbol{\beta}$. Again, parameters $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ are independent *a posteriori* conditional on imputed $\boldsymbol{S}$. As such, HMC is performed independently for these two sets of parameters to ensure numerical stability: it is crucial to calculate the Cholesky decomposition of the mass matrix when performing the leapfrog steps, and splitting $\boldsymbol{\theta}$ into the two

sets helps prevent numerical errors. Another model considered by Gill et al. (2013) uses exclusion restriction by assuming $\beta_{n00} = \beta_{n10}$. We apply HMC algorithm to this model as well, with the change that $L_\beta = 75$.

C.3 Data and Models for Example 2

Table C.2 summarizes the observed data used for Example 2 in Section 4.4.3.

Table C.2: Observed outcomes in the Hirano et al. (2000) analysis.

Encouragement z_i^{obs}	Vaccination d_i^{obs}	Hospitalization c_i^{obs}	# Subjects	Strata
0	0	0	1040	c or n
0	0	1	99	c or n
0	1	0	237	a
0	1	1	30	a
1	0	0	944	n
1	0	1	85	n
1	1	0	424	c or a
1	1	1	31	c or a

A summary of our model assumptions is below.

$$y_i(0) \perp\!\!\!\perp y_i(1) \mid \boldsymbol{x}_i, \boldsymbol{\theta},$$

$$z_i \mid y_i(0), y_i(1), d_i(0), d_i(1), s_i, \boldsymbol{x}_i, \boldsymbol{\theta} \;\sim\; z_i \mid \boldsymbol{x}_i, \boldsymbol{\theta},$$

$$\Psi_c(\boldsymbol{x}_i) = \Pr(s_i = c \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp(\psi_{c0} + \psi_{c1}x_{i1} + \psi_{c2}x_{i2})}{1 + \exp(\psi_{c0} + \psi_{c1}x_{i1} + \psi_{c2}x_{i2}) + \exp(\psi_{a0} + \psi_{a1}x_{i1} + \psi_{a2}x_{i2})},$$

$$\Psi_a(\boldsymbol{x}_i) = \Pr(s_i = a \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp(\psi_{a0} + \psi_{a1}x_{i1} + \psi_{a2}x_{i2})}{1 + \exp(\psi_{c0} + \psi_{c1}x_{i1} + \psi_{c2}x_{i2}) + \exp(\psi_{a0} + \psi_{a1}x_{i1} + \psi_{a2}x_{i2})},$$

$$\Gamma(\boldsymbol{x}_i, \beta_{sz0}) = \Pr\left\{y_i(z_i) = 1 \mid s_i = s, z_i = z, \boldsymbol{x}_i, \boldsymbol{\theta}\right\} = \frac{\exp(\beta_{sz0} + \beta_{0{\cdot}1} x_{i1} + \beta_{0{\cdot}2} x_{i2} + \beta_{0{\cdot}3} x_{i3} + \beta_{0{\cdot}4} x_{i4})}{1 + \exp(\beta_{sz0} + \beta_{0{\cdot}1} x_{i1} + \beta_{0{\cdot}2} x_{i2} + \beta_{0{\cdot}3} x_{i3} + \beta_{0{\cdot}4} x_{i4})},$$

$$\boldsymbol{\theta} = (\psi_{c0}, \psi_{c1}, \psi_{c2}, \psi_{a0}, \psi_{a1}, \psi_{a2}, \beta_{c00}, \beta_{c10}, \beta_{n00}, \beta_{n10}, \beta_{a00}, \beta_{a10}, \beta_{0{\cdot}1}, \beta_{0{\cdot}2}).$$

Again, intercepts and slopes follow Cauchy distribution with scale 2.5 independently *a priori*, and continuous covariates are standardized to have mean 0 and standard deviation 2.5. The complete-data likelihood follows below.

$$\begin{aligned}
&\prod_{i:s_i=c} \Bigg[\prod_{i\in(0,0)} \left\{\Psi_c(\boldsymbol{x}_i) \times \Gamma(\boldsymbol{x}_i, \beta_{c00})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{c00})\right)^{1-y_i}\right\} \\
&\qquad \prod_{i\in(1,1)} \left\{\Psi_c(\boldsymbol{x}_i) \times \Gamma(\boldsymbol{x}_i, \beta_{c10})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{c10})\right)^{1-y_i}\right\} \Bigg] \times \\
&\prod_{i:s_i=a} \Bigg[\prod_{i\in(0,1)} \left\{\Psi_a(\boldsymbol{x}_i) \times \Gamma(\boldsymbol{x}_i, \beta_{a00})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{a00})\right)^{1-y_i}\right\} \times \\
&\qquad \prod_{i\in(1,0)} \left\{\Psi_a(\boldsymbol{x}_i) \times \Gamma(\boldsymbol{x}_i, \beta_{a10})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{a10})\right)^{1-y_i}\right\} \Bigg] \\
&\prod_{i:s_i=n} \Bigg[\prod_{i\in(0,0)} \left\{(1 - \Psi_c(\boldsymbol{x}_i) - \Psi_a(\boldsymbol{x}_i)) \times \Gamma(\boldsymbol{x}_i, \beta_{n00})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{n00})\right)^{1-y_i}\right\} \times \\
&\qquad \prod_{i\in(1,0)} \left\{(1 - \Psi_c(\boldsymbol{x}_i) - \Psi_a(\boldsymbol{x}_i)) \times \Gamma(\boldsymbol{x}_i, \beta_{n10})^{y_i} \left(1 - \Gamma(\boldsymbol{x}_i, \beta_{n10})\right)^{1-y_i}\right\} \Bigg]
\end{aligned}$$

Note that there are four different models under consideration, depending on whether the exclusion restriction is placed on never-takers ($\beta_{n00} = \beta_{n10}$) or always-takers ($\beta_{a00} = \beta_{a10}$). Figures C.1 and C.2 show autocorrelation plots of draws generated for the case with no exclusion restrictions.

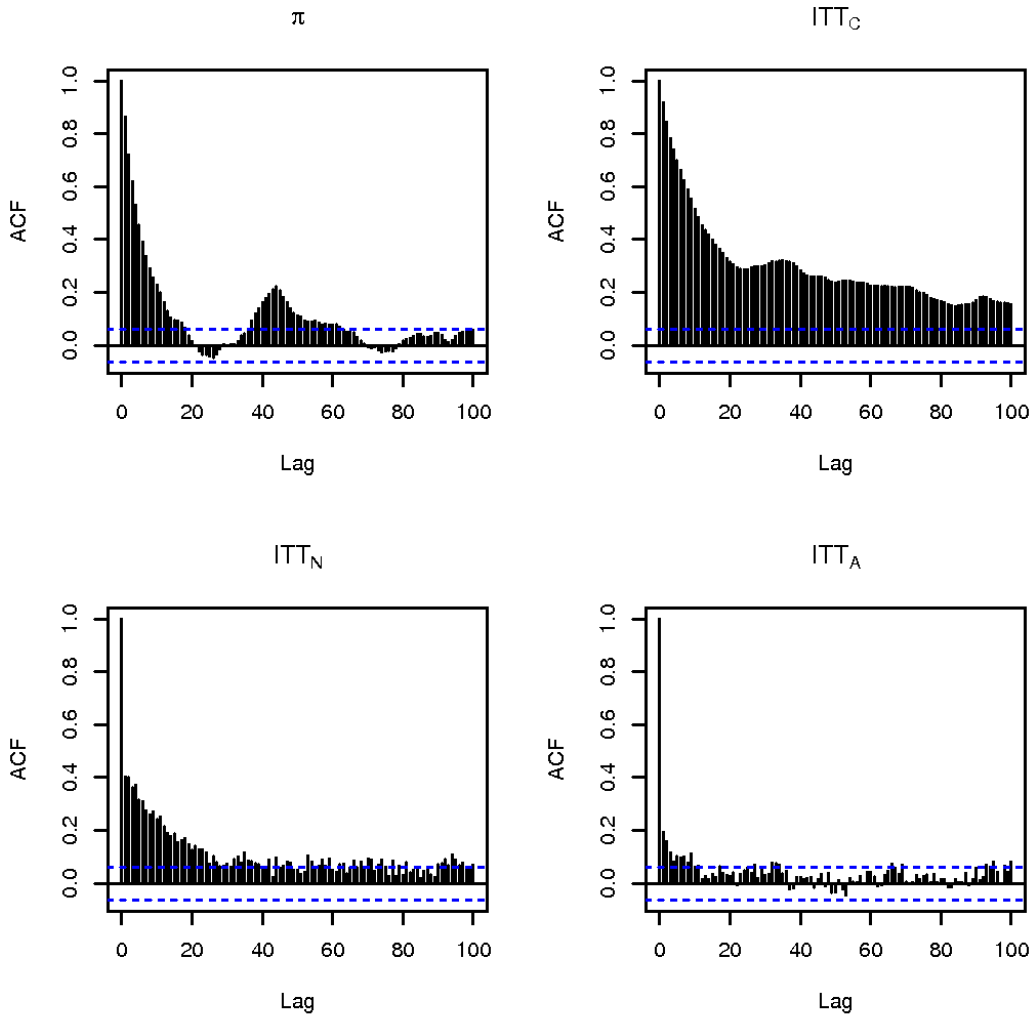


Figure C.1: Autocorrelation plots of draws produced by HMC algorithm in Example 2, described in Section 4.4.3, for the case with exclusion restrictions.

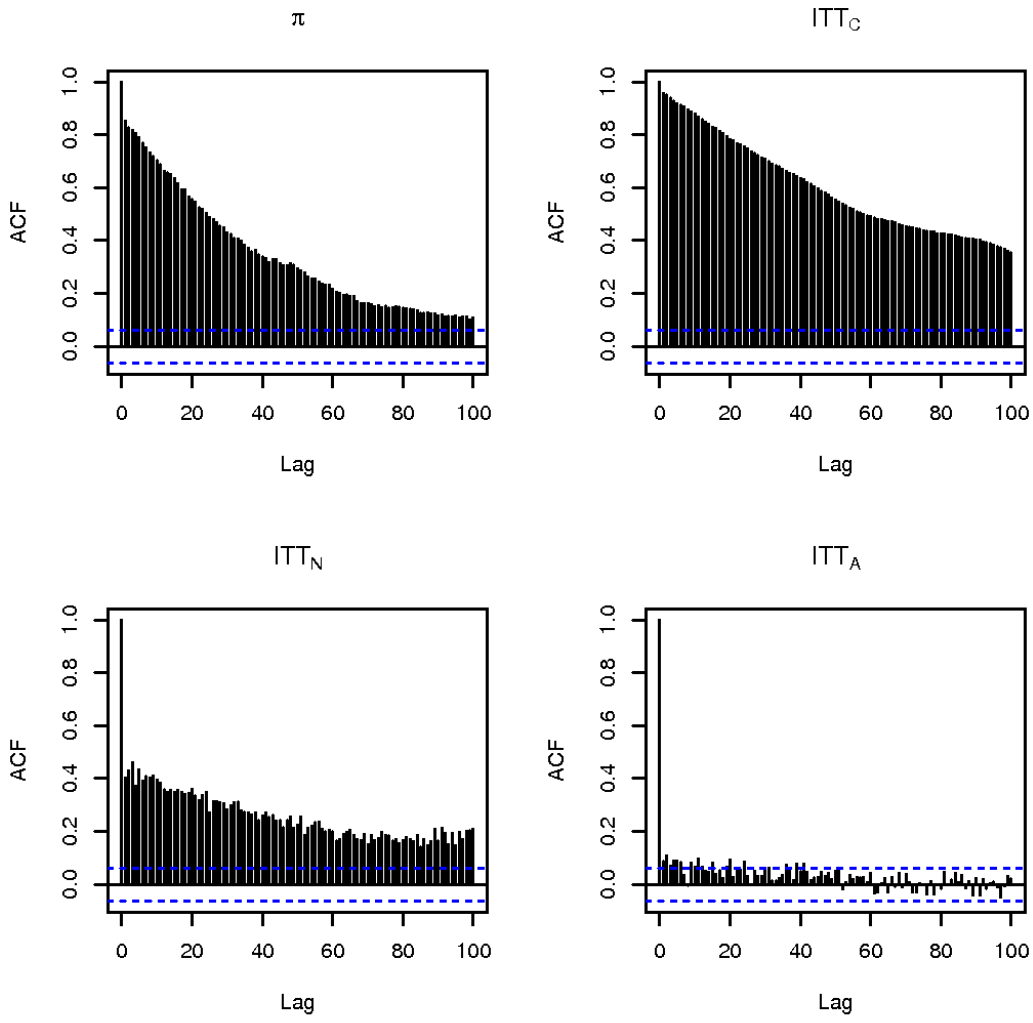


Figure C.2: Autocorrelation plots of draws produced by MHMC algorithm in Example 2, described in Section 4.4.3, for the case with exclusion restrictions. Again, when comparing to the plots in Figure C.1, it is evident that the correlations between consecutive and near-consecutive draws are much higher for MHMC.

Bibliography

Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade, M., Allman, R. M., Meleth, S., and Bourge, R. C. (2006). "Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods." *European Heart Journal*, 27(12):1431–1439.

Allan, F. E. and Wishart, J. (1930). "A Method of Estimating the Yield of a Missing Plot in Field Experimental Work." *The Journal of Agricultural Science*, 20(2):399–406.

Allison, P. D. (2001) *Missing Data*. Sage Publications, Inc, 1 edition.

Aylward, B. S., Anderson, R. A., and Nelson, T. D. (2010). "Approaches to handling missing data within developmental and behavioral pediatric research." *Journal of developmental and behavioral pediatrics: JDBP*, 31(1):54–60.

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2003). "Principal Stratification Approach to Broken Randomized Experiments." *Journal of the American Statistical Association*, 98(462):299–323.

Barnard, J. and Rubin, D. B. (1999). "Small-Sample Degrees of Freedom with Multiple Imputation." *Biometrika*, 86(4):948–955.

Bodner, T. E. (2006). "Missing data: prevalence and reporting practices." *Psychological Reports*, 99(3):675–680.

Burda, M. and Maheu, J. (2011). "Bayesian Adaptive Hamiltonian Monte Carlo with an Application to High-Dimensional BEKK GARCH Models." Technical report.

Burton, A. and Altman, D. G. (2004). "Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines." *British Journal of Cancer*, 91(1):4–8.

Burzykowski, T., Carpenter, J., Coens, C., Evans, D., France, L., Kenward, M., Lane, P., Matcham, J., Morgan, D., Phillips, A., Roger, J., Sullivan, B., White, I., and

Yu, L.-M. (2010). "Missing data: discussion points from the PSI missing data expert group." *Pharmaceutical Statistics*, 9(4):288–297.

Buuren, S. v. (2012) *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, 1 edition.

Buuren van, S. and Groothuis-Oudshoorn, K. (2011). "MICE: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3).

Cacoullos, T. (1965a). "Comparing Mahalanobis Distances I: Comparing Distances between k Known Normal Populations and Another Unknown." *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 27(1):1–22.

— (1965b). "Comparing Mahalanobis Distances II: Bayes Procedures When the Mean Vectors Are Unknown." *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 27(1):23–32.

Campbell, G., Pennello, G., and Yue, L. (2011). "Missing Data in the Regulation of Medical Devices." *Journal of Biopharmaceutical Statistics*, 21(2):180–195.

Carpenter, J. R. and Kenward, M. G. (2008). "Missing data in clinical trials - a practical guide." *National Institute for Health Research, Publication RM03/JH17/MK: Birmingham*.

CHMP (2010). "Guideline on Missing Data in Confirmatory Clinical Trials." Technical report, European Medical Agency.

Cox, D. R. and Wermuth, N. (1993). "Linear Dependencies Represented by Chain Graphs." *Statistical Science*, 8(3):204–218.

D'Agostino Jr., R. B. (1998). "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." *Statistics in Medicine*, 17(19):2265–2281.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). "Review: a gentle introduction to imputation of missing values." *Journal of Clinical Epidemiology*, 59(10):1087–1091.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, 195(2):216–222.

Egleston, B. L., Cropsey, K. L., Lazev, A. B., and Heckman, C. J. (2010). "A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies." *Clinical Trials*, 7(3):286–298.

Elliott, M. R., Raghunathan, T. E., and Li, Y. (2010). "Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes." *Biostatistics*, 11(2):353–372.

Frangakis, C. E. and Rubin, D. B. (1999). "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes." *Biometrika*, 86(2):365–379.

— (2002). "Principal Stratification in Causal Inference." *Biometrics*, 58(1):21–29.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data." *Journal of the American Statistical Association*, 107(498):450–466.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., and Ten Have, T. R. (2009). "Mediation analysis with principal stratification." *Statistics in Medicine*, 28(7):1108–1130.

Gelman, A. and Rubin, D. B. (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, 7(4):457–472.

Gerber, A. and Green, D. (2000). "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American Political Science Review*, 94(3):653–663.

Gerber, A. S. and Green, D. P. (2005). "Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)." *American Political Science Review*, 99(02).

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). "Sensitivity Analysis for the Assessment of Causal Vaccine Effects on Viral Load in HIV Vaccine Trials." *Biometrics*, 59(3):531–541.

Gill, M., Sabbaghi, A., and Schneer, B. (2013). "Identification of the Causal Effect of Canvassing using Instrumental Variables." Technical report.

Girolami, M. and Calderhead, B. (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). "Selection modeling versus mixture modeling with nonignorable nonresponse." In *Drawing inferences from self-selected samples*, 115–142. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Greenland, S. and Finkle, W. D. (1995). "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses." *American Journal of Epidemiology*, 142(12):1255–1264.

Hansen, B. B. and Bowers, J. (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, 23(2):219–236.

Heckman, J. J. (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." NBER chapters, National Bureau of Economic Research, Inc.

Heitjan, D. F. and Basu, S. (1996). "Distinguishing "Missing at Random" and "Missing Completely at Random"." *The American Statistician*, 50(3):207–213.

Held, L. (2004). "Simultaneous Posterior Probability Statements from Monte Carlo Output." *Journal of Computational and Graphical Statistics*, 13(1):20–35.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000). "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics*, 1(1):69–88.

Hockney, R. W. (1970). "Potential Calculation and Some Applications." *Methods Comput. Phys. 9: 135-211(1970)*..

Holland, P. W. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396):945–960.

Hollis, S. (2002). "A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome." *Statistics in Medicine*, 21(24):3823–3834.

Horvitz, D. and Thompson, D. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47(260):663–685.

Hotelling, H. (1931). "The Generalization of Student's Ratio." *The Annals of Mathematical Statistics*, 2(3):360–378.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003). "On the analysis of viral load endpoints in HIV vaccine trials." *Statistics in Medicine*, 22(14):2281–2298.

Imbens, G. W. and Angrist, J. D. (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (1997). "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics*, 25(1):305–327.

Jamshidian, M. and Jalal, S. (2010). "Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data." *Psychometrika*, 75(4):649–674.

Jelicic, H., Phelps, E., and Lerner, R. M. (2009). "Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology." *Developmental Psychology*, 45(4):1195–1199.

Jin, H. and Rubin, D. B. (2009). "Public Schools Versus Private Schools: Causal Inference With Partial Compliance." *Journal of Educational and Behavioral Statistics*, 34(1):24–45.

Kim, K. and Bentler, P. (2002). "Tests of homogeneity of means and covariance matrices for multivariate incomplete data." *Psychometrika*, 67(4):609–623.

Klebanoff, M. A. and Cole, S. R. (2008). "Use of Multiple Imputation in the Epidemiologic Literature." *American Journal of Epidemiology*, 168(4):355–357.

Lee, S.-Y. (2007) *Handbook of latent variable and related models*. Elsevier.

Licht, C. (2010). "New methods for generating significance levels from multiply-imputed data." Ph.D. thesis, Otto-Friedrich-Universitat.

Little, R. J. A. (1986). "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review / Revue Internationale de Statistique*, 54(2):139–157.

— (1988a). "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics*, 6(3):287–296.

— (1988b). "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association*, 83(404):1198–1202.

— (1992). "Regression With Missing X's: A Review." *Journal of the American Statistical Association*, 87(420):1227–1237.

— (1994). "A Class of Pattern-Mixture Models for Normal Incomplete Data." *Biometrika*, 81(3):471–483.

— (1995). "Modeling the Drop-Out Mechanism in Repeated-Measures Studies." *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical analysis with missing data*. Wiley, 1 ed. edition.

— (2002) *Statistical analysis with missing data*. Wiley, 2nd. edition.

Liu, J. S. (2008) *Monte Carlo Strategies in Scientific Computing*. Springer.

Liublinska, V. and Rubin, D. B. (2012). "Re: "Dealing With Missing Outcome Data in Randomized Trials and Observational Studies"." *American Journal of Epidemiology*, 176(4):357–358.

Mackinnon, A. (2010). "The use and reporting of multiple imputation in medical research a review." *Journal of Internal Medicine*, 268(6):586–593.

Matts, J. P., Launer, C. A., Nelson, E. T., Miller, C., and Dain, B. (1997). "A graphical assessment of the potential impact of losses to follow-up on the validity of study results." *Statistics in Medicine*, 16(17):1943–1954.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models, Second Edition*. CRC Press.

McKnight, P. E. (2007) *Missing data: a gentle introduction*. Guilford Press.

Mealli, F. and Rubin, D. B. (2013). "Missing at Random for Independent and Identically Distributed Variables. (In progress)."

Meng, X. L. (1997). "The EM algorithm and medical studies: a historical link." *Statistical Methods in Medical Research*, 6(1):3–23.

M'Kendrick, A. G. (1925). "Applications of Mathematics to Medical Problems." *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Molenberghs, G., Goetghebeur, E. J. T., Lipsitz, S. R., and Kenward, M. G. (1999). "Nonrandom Missingness in Categorical Data: Strengths and Limitations." *The American Statistician*, 53(2):110–118.

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., and Harrell, J., Frank E (2006). "Using the outcome for imputation of missing predictor values was preferred." *Journal of Clinical Epidemiology*, 59(10):1092–1101.

Morris, C. N. (1982). "Natural Exponential Families with Quadratic Variance Functions." *The Annals of Statistics*, 10(1):65–80.

— (1983). "Natural Exponential Families with Quadratic Variance Functions: Statistical Theory." *The Annals of Statistics*, 11(2):515–529.

Morris, C. N. and Lock, K. F. (2009). "Unifying the Named Natural Exponential Families and Their Relatives." *The American Statistician*, 63(3):247–253.

Neal, R. M. (1995). "Bayesian Learning for Neural Networks." Ph.D. thesis, University of Toronto.

— (2011). "MCMC using Hamiltonian dynamics." In *Handbook of Markov Chain Monte Carlo*, 113–162. Chapman and Hall/CRC, 1 edition.

Newgard, C. D. and Haukoos, J. S. (2008). "Advanced Statistics: Missing Data in Clinical ResearchPart 2: Multiple Imputation." *Academic Emergency Medicine*, 14(7):669–678.

NRC-Panel (2010) *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press.

Pearl, J. (2009) *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Peugh, J. L. and Enders, C. K. (2004). "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement." *Review of Educational Research*, 74(4):525 –556.

Raghunathan, T., Solenberger, P., and Van Hoewyk, J. (2002). "IVEware: Imputation and Variance Estimation Software." Technical report.

Resseguier, N., Giorgi, R., and Paoletti, X. (2011). "Sensitivity Analysis When Data Are Missing Not-at-random." *Epidemiology*, 22(2):282.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association*, 89(427):846–866.

— (1995). "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. R. and Rubin, D. B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1):41–55.

— (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5):688–701.

— (1976). "Inference and missing data." *Biometrika*, 63(3):581 –592.

— (1977). "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association*, 72(359):538–543.

— (1980). "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association*, 75(371):591–593.

— (1986). "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business & Economic Statistics*, 4(1):87–94.

— (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1st edition.

— (1990). "Formal Modes of Statistical Inference For Causal Effects." *Journal of Statistical Planning and Inference*, 25:279–292.

— (1998). "More powerful randomization-based p-values in double-blind trials with non-compliance." *Statistics in Medicine*, 17(3):371–385.

— (2003). "Nested multiple imputation of NMES via partially incompatible MCMC." *Statistica Neerlandica*, 57(1):318.

— (2004) *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, 2nd edition.

— (2006a). "Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death." *Statistical Science*, 21(3):299–309.

— (2006b) *Matched Sampling for Causal Effects*. Cambridge University Press.

— (2007). "Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics." In C.R. Rao, J. M. and Rao, D. (eds.), *Handbook of Statistics*, volume Volume 27, 28–63. Elsevier.

Rubin, D. B. and Schenker, N. (1991). "Multiple imputation in healthcare databases: An overview and some applications." *Statistics in Medicine*, 10(4):585–598.

Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. CRC Press.

— (1999). "Multiple imputation: a primer." *Statistical Methods in Medical Research*, 8(1):3–15.

Schafer, J. L. and Graham, J. W. (2002). "Missing data: our view of the state of the art." *Psychological Methods*, 7(2):147–177.

Schulz, K. F., Altman, D. G., and Moher, D. (2010). "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials." *Annals of Internal Medicine*.

Shepherd, B. E., Gilbert, P. B., and Lumley, T. (2007). "Sensitivity Analyses Comparing Time-to-Event Outcomes Existing Only in a Subset Selected Postrandomization." *Journal of the American Statistical Association*, 102(478):573–582.

Shepherd, B. E., Redman, M. W., and Ankerst, D. P. (2008). "Does Finasteride Affect the Severity of Prostate Cancer? A Causal Sensitivity Analysis." *Journal of the American Statistical Association*, 103(484):1392–1404.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82(398):528–540.

Templ, M. and Filzmoser, P. (2008). "Visualization of Missing Values using the R-Package VIM." Technical Report CS-2008-1, Vienna University of Technology, Vienna, Austria.

van der Heijden, G. J., T. Donders, A. R., Stijnen, T., and Moons, K. G. (2006). "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example." *Journal of Clinical Epidemiology*, 59(10):1102–1109.

Wainer, H. (ed.) (1986) *Drawing inferences from self-selected samples*, volume xii. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Weatherall, M., Pickering, R., and Harris, S. (2009). "Graphical Sensitivity Analysis with Different Methods of Imputation for a Trial with Probable Non-Ignorable Missing Data." *Australian & New Zealand Journal of Statistics*, 51(4):397–413.

Welch, B. L. (1938). "The Significance of the Difference Between Two Means when the Population Variances are Unequal." *Biometrika*, 29(3/4):350–362.

White, I. R., Royston, P., and Wood, A. M. (2011). "Multiple imputation using chained equations: Issues and guidance for practice." *Statistics in Medicine*, 30(4):377–399.

Wilkinson, L. (1999). "Statistical methods in psychology journals: Guidelines and explanations." *American Psychologist*, 54(8):594–604.

Wilks, S. S. (1932). "Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples." *The Annals of Mathematical Statistics*, 3(3):163–195.

Yan, X., Lee, S., and Li, N. (2009). "Missing Data Handling Methods in Medical Device Clinical Trials." *Journal of Biopharmaceutical Statistics*, 19(6):1085–1098.

Yates, F. (1933). "The analysis of replicated experiments when the field results are incomplete." *Empire Journal of Experimental Agriculture*, 1(3):129–42.

Zhang, J. L. and Rubin, D. B. (2003). "Estimation of Causal Effects Via Principal Stratification When Some Outcomes Are Truncated by "Death"." *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2008). "Evaluating the effects of job training programs on wages through principal stratification." *Advances in Econometrics*, 21:117–145.

— (2009). "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification." *Journal of the American Statistical Association*, 104(485):166–176.